LECTURES ON GRAVITATION
AND RELATIVITY

Goddard Institute for Space Studies
New York, New York

Volume II

THE SIGNIFICANCE FOR THE SOLAR SYSTEM OF TIME VARYING GRAVITATION

N68 - 14305

Lecture VIII

R. H. Dicke

Princeton University

Seminar on Gravitation and Relativity, NASA Goddard Space Flight Center, Institute for Space Studies, New York, N.Y.; edited by H. Y. Chiu and W. F. Hoffmann.

Preface

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R. H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution the space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a minimum amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and correct the final manuscript. Hence, we accept responsibility for errors and omissions.

H. Y. Chiu

W. F. Hoffmann

In my last lecture, I discussed a number of general features of gravitation theory. Two of these features furnish the starting point for this lecture.

1) The requirements that a theory of gravitation be expressed in generally covariant equations and that inertial and gravitational forces both be obtained from a single invariant lead us naturally to represent gravitational effects by a tensor field. Einstein's theory is a particular example of a tensor theory in which the tensor field is the only field exhibiting gravitational effects and the geometry is so defined that this tensor is the metric tensor of the geometry.

2) Einstein's general theory of relativity is not relativistic in the Machian sense. That is, this theory is not limited to a description of the relations between positions of matter. Rather, properties such as fixed directions are ascribed to empty space in the complete absence of matter, and motion is referred to a preferred, or absolute, geometry.

In this lecture, I will discuss how Einstein's theory can be modified to overcome in part its absolute space-time character by introducing a second field quantity into the equations. Then, I will go on to relate this modification to some of the problems connected with geophysics and astrophysics.

I anticipated some of the characteristics of this modification in the last lecture with the discussion of Sciama's model. Sciama and others have pointed out that it is possible to incorporate Mach's principle more completely into general relativity by introducing a gravitational constant which is not strictly a constant but is a function of some field variable.

This possibility has also been raised in connection with a time varying gravitational constant. The suggestion of a time varying gravitational constant may have first appeared in physics in connection with Milne's ideas of cosmology. Later Dirac (1) suggested that certain numerical coincidences of large cosmological numbers might imply time-varying gravitation, that is, a gravitational constant which is not truly a constant, but a function of time. Later Jordan (2) attempted to put Dirac's ideas into a proper field-theoretic form by introducing a gravitational "constant," a variable "constant" as function of a scalar.

I shall approach this matter in a slightly different way, by pointing out explicitly that in addition to the gravitational field which is connected with the geometry of space, one can

---

(1) P. A. M. Dirac, Proceedings Royal Society (London) A 165, 199 (1938).
(2) P. Jordan, Schwerkraft und Weltall (Vieweg und Sohn, Braunschweig, 1955).

have long-range matter-type fields. We are already familiar with one of these, the electromagnetic field. Two charged particles can interact with each other over a long distance through electromagnetic fields. In the previous lecture I showed that this kind of vector field appears to be ruled out for producing long range effects of cosmological importance. Large sections of the universe cannot interact with each other through vector interactions obeying Maxwell-type (gauge invariant) equations in a uniformly isotropic universe.

Within the framework of relativity there are two other fields that might play important roles in cosmology. One of these is a scalar long-range field, and the other is a tensor long-range field, that is, a second tensor field, in addition to the metric tensor that is associated with gravitation in Einstein's theory. I shall take these in inverse order, discussing the tensor field first.

It is possible to have a tensor interaction in addition to the metric tensor of space. But it would be very difficult to incorporate into the theory another such long-range tensor field. A second tensor field would be expected to lead to some queer results, results that would show up in experiments of the kind that Hughes and his students performed.

Professor Hughes, and independently Drever, was able to demonstrate the isotropy of space to a very high precision. This experiment is discussed in detail in Lecture 6. The reason the tensor field is expected to run into difficulty with the Hughes experiment was first discussed quantitatively by J. Peebles.[(2a)] It is this: Assume there is some second tensor field. We may choose a coordinate system for which the metric tensor is locally Minkowskian. Then, generally speaking, the second tensor will not be locally Minkowskian but will have some form for which the spatial parts of this tensor exhibit an anisotropy. If there are forces associated with this second tensor field, it would be expected that this spatial anisotropy would appear in the results of the Hughes experiment.

If we should happen to find isotropy simultaneously for both tensors in a particular coordinate system, then we could simply transform to a moving coordinate system. If the tensors are not identical, we can always, by moving, obtain a lack of spatial isotropy in one of the tensors. I believe that, because of the tremendous precision and sensitivity of the experiment, this is a compelling argument against the second tensor interaction. I see no very obvious way of getting a second long-range tensor

---

(2a) J. Peebles & R. H. Dicke, "The Significance of Spatial Isotropy" (to be published).

field into physics.

The case for the scalar field is more promising. I shall summarize its properties which were mentioned in my last lecture. It is remarkable that for the little we know about this interaction in an observational way (in fact, even its existence is in doubt) we can delineate its properties so well. This is because the interaction is so simple that a couple of observations and the requirement of Lorentz invariance are sufficient to specify fairly completely its properties. The properties of a long-range scalar interaction (a neutral, scalar, massless field) are the following:

1) The scalar field can only be weak. Its strength must be of the order of the gravitational interaction.

2) The interaction of a scalar field with a particle cannot occur unless the mass of the particle is a function of the scalar.

This means that work must be done on the particle to move it in a non-constant scalar field. This implies an extra force which acts on matter by virtue of its interaction with the scalar. We express this in the form of an equation by

$$\frac{d}{d\tau}(mu_i) - \frac{1}{2}mg_{jk,i}u^j u^k + m_{,i} = 0 \qquad (1)$$

where

$$u^i u_i = -1 \tag{2}$$

and

$$m_{,i} = \frac{dm}{d\varphi} \varphi_{,i} \tag{3}$$

The great accuracy of the Eötvös experiment imposes the requirement that if there is such a scalar interaction, all particles must suffer essentially the same type of scalar interaction. Otherwise, there would be anomalous accelerations. In other words, if the mass of the proton plus electron varied with the scalar with some different functional dependence than the mass of the neutron, then, generally speaking, a neutron would fall with a different acceleration from that of an ordinary hydrogen atom. So we face the requirement that the functional dependence of the mass on the scalar field variable should be equal to some constant times some standard function which is the same for all particles.

$$m(\varphi) = m_0 f(\varphi) \tag{4}$$

The assumption that the mass of a particle is a function of a scalar leads to some rather strange effects. In particular, the gravitational coupling "constant," then, is not really constant. The gravitational constant can be given as a dimensionless number, a coupling constant, in terms of atomic constants as

$$\frac{Gm_p^2}{\hbar c} \simeq 10^{-40} \tag{5}$$

where m_p is the mass of the proton. But if the mass of the proton varies from one place to the other, this ratio also will vary so that the gravitational interaction expressed in atomic units is not constant.

Normally the dimensionless coupling constants of physics, such as

$$\alpha = \frac{e^2}{\hbar c} \simeq \frac{1}{137} \tag{6}$$

and

$$\frac{Gm_p^2}{\hbar c} \sim 10^{-40} \tag{7}$$

are regarded as fait accompli of nature, numbers preordained and unrelated to other physical dimensionless numbers. However, if the gravitational coupling "constant" is variable, and is determined as a function of some scalar field variable, in turn determined by the matter distribution in the universe, it becomes possible to understand the extraordinary value of this number. In my opinion, the number 10^{-40}, contrary to what Eddington thought, would not be expected to appear in a formalism as a pure number of simple mathematical origin. However, with the above interpretation, $\frac{G m_p^2}{\hbar c}$ may be considered to be small because the universe has so many particles ($\sim 10^{80}$). This large amount of matter, at great distances in the universe, generates a local value of m_p such that the gravitational coupling constant is small.

There is another remarkable feature associated with varying particle masses. If the masses of atoms vary, so also do their periods and diameters and as a result the lengths of meter sticks and the periods of clocks. These are all affected by the value of the scalar field. A meter stick at one place has a different length than a meter stick somewhere else (Figure 1) because the geometry which I have been using to describe these effects is not the geometry which is measured by meter sticks and clocks but that given by the metric tensor of Einstein's equations.

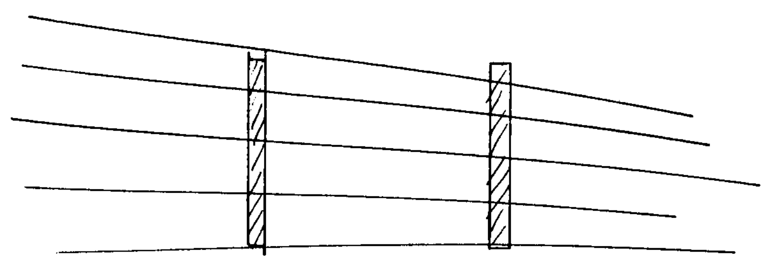


FIGURE 1. The unevenly spaced lines represent the curved geometry given by Einstein equations relative to which the length of a meter stick varies from place to place as a function of the scalar field.

We have a problem then of redefining the geometry, if we wish, in such a way that the length of a meter stick does not change. That is, we may define the unit of length everywhere to be that length measured by a meter stick which is transported from place to place (Figure 2). If I redefine the geometry in this way, such that meter sticks and clocks do behave

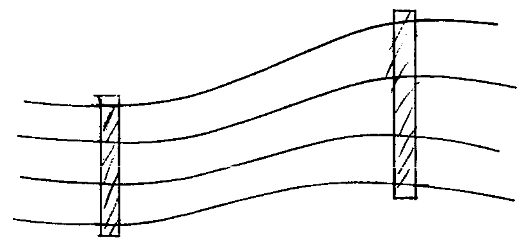


FIGURE 2. Relative to an appropriately redefined geometry not satisfying Einstein's equations the length of a meter stick is constant.

properly, and the masses of particles do not vary, then I discover that two things happen:

1) The field equation for the metric tensor is not that of general relativity. It is a modified equation.

2) With these modified equations, the gravitational constant is not a constant, but varies from one place to another. All the other physical constants are properly constant.

This type of formalism for which the equations of general relativity are replaced by some modified equations was first introduced by Jordan. The particular modification I will use is closely related to one of the forms of Jordan's equations.

The most compact way of presenting the theory is in terms of the variational principles from which the equations are obtained. In general relativity one gets the Einstein field equations and equations for the motion of matter from a variational principle of this kind:

$$\delta \int (R + GL)\sqrt{-g}\, d^4x = 0 \tag{8}$$

R is the contracted curvature tensor, G is the Gravitational constant, and L is the Lagrangian density of matter. If we carry out the indicated variation for the metric tensor components we obtain Einstein's field equations. If we carry out a variation on the particle coordinates that appear in the Lagrangian density for matter, we obtain the equations of motion of matter. All the equations of physics are contained in the variational principle.

In order to introduce a scalar field explicitly, I must add to the Lagrangian density of matter a Lagrangian density for the scalar field. The variational principle then has the form:

$$\delta \int [R + G(L + L_\lambda)]\sqrt{-g}\, d^4x = 0 \qquad (9)$$

where for convenience we choose L_λ to have the form

$$L_\lambda = -G^{-1}(\omega + 3/2)\frac{\lambda_{,i}\lambda^{,i}}{\lambda^2} \qquad (10)$$

λ is the scalar field. ω is a constant which can be thought of as a coupling constant for the field. In addition to the scalar appearing explicitly in the scalar Lagrangian density it appears also explicitly in the masses of the particles in the matter Lagrangian. This is assumed to be of the form

$$m = m_\circ \lambda^{-1/2} \qquad (11)$$

I will indicate later why this form is particularly interesting.

From equation (9) we obtain the Einstein field equations for the components of the metric tensor and new equations of motion for particles. In the units given by this geometry meter sticks behave strangely. They contract and expand as they are moved from one place to another. Clocks run fast one place and slow another. But we can redefine the units of measure in such a way that meter sticks do not have this strange behavior. We discover that the corresponding transformation of the equations leads to the following variational principle where φ is a new scalar with dimensions of G^{-1} and $\varphi \sim \lambda$.

$$\delta \int \left[\varphi R - \frac{\omega \varphi_{,i} \varphi^{,i}}{\varphi} + L \right] \sqrt{-g} \, d^4x = 0 \qquad (12)$$

In this equation the particle masses appearing in L are no longer variable but are now constant.

This is similar to one of the Jordan-type variational principles. It leads to gravitational interactions which are described not by a metric tensor, but a metric tensor plus a scalar. Equation (5) is essentially one of the Jordan equations. It was first discussed in relation to Mach's Principle by C. Brams and R. H. Dicke (3). The transformation from equation (12) to equation (9) was discussed in (4). Matter obeys the

---

(3) C. Brans & R. H. Dicke, Physical Review, **124**, 925 (1961).

(4) R. H. Dicke, Physical Review, **125**, 2163, (1962)

usual kinds of equations of motion that we are familiar with from Einstein's theory. But Einstein's field equations are not valid.

To summarize, in the form of the theory in which Einstein's equations are satisfied meter sticks behave in a strange way. In the form in which meter sticks behave properly, the Einstein field equations are not valid. We have the choice of one or the other form. They are completely equivalent physical descriptions. They differ from each other only in the way that we have defined our units of measure and hence our geometry. This equivalence has been discussed in (3).

It is interesting that equations of this kind seem to be compatible with Mach's principle. Also they imply a position and time varying gravitational constant. It appears difficult to obtain a theory incorporating Mach's principle with only a metric tensor and without a scalar field. I shall indicate what this type of theory implies about the variation of gravity with time and position. Then I would like to describe as nearly as I can what would be the effect on the solar system of the gravitational interaction changing with time.

For a static situation, and from equation (12), the field equation for the scalar field is

$$\nabla^2\varphi \sim T \tag{13}$$

where T is the contracted energy momentum tensor of matter. For a time-varying matter configuration we must replace the Laplacian operator by the d'Alembertian operator. In a Minkowskian coordinate system, neglecting curvature effects, this operator has the form shown in equation (14):

$$\Box^2\varphi \equiv \nabla^2\varphi - \frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} \simeq T \tag{14}$$

This equation implies that as the universe, assumed uniform, expands with time and the amount of visible matter in the universe changes, the scalar connected with that matter distribution changes.

As a consequence gravitation must get weaker with time. One can make a quite reasonable guess as to how fast it should get weaker with time. This turns out to be of the order of 3 parts in 10^{11} per year, assuming the theory as presented here is valid.

This variation implies many interesting observable effects concerning the history and present state of the solar system and the galaxy. However, the data that one has to deal with always haveso many possible explanations that there is little if any hope of demonstrating from the following considerations that a variation in G occurs. But on the other hand, if you were to give me a good laboratory proof that the gravitational constant is changing with time, then my remarks concerning expected effects which would occur in connection with this changing G, would not be unreasonable.

One of the primary effects of a gravitational constant getting smaller with time is connected with stellar evolution. The reason for this is that the luminosity of a star is a rather sensitive function of the gravitational constant. If the gravitational constant were getting weaker with time the luminosity of a star would be decreasing. This would affect our observations in two different ways:

1) The lifetime of a star is now determined from its present state of evolution assuming a constant gravitation constant in the past. This determination would be wrong if the star had evolved more rapidly in the past because of a stronger gravitational constant.

2) The sun would have had a greater luminosity in the past than it has now. This would lead to higher surface temperatures of the earth and other planets in the past.

I will examine the effect of a changing G on stellar evolution first. One can describe the situation rather simply. As a consequence of the virial theorem, the gravitational potential energy of a star is of the order of magnitude of the kinetic energy. This in turn is proportional to the central temperature of the star because the central temperature is a measure of the kinetic average energy of the particles. Therefore we may write

$$\frac{GM}{R} \sim T \tag{15}$$

where M, R and T are the mass, radius, and central temperature of the star respectively. If we hold the radius constant while allowing the gravitational constant to change we can see that the change in central temperature is proportional to the change in G. On the other hand, the rate at which a black body radiates is proportional to the fourth power of the temperature. Therefore, in the most naive way, we would expect the luminosity to vary as the fourth power of the gravitational constant.

The situation is not quite this simple. First of all, this argument would only be true if the opacity were temperature-independent. In the case of very massive stars (very bright stars) where Compton scattering plays a dominant role in determining the opacity, the Compton cross-section of electron is fixed and is independent of the temperature. In that case we do expect the luminosity to vary roughly as the fourth power of the gravitational constant.

$$L \propto G^4 \qquad \text{(for very massive star)} \tag{16}$$

However, in the case of a star of the order of the sun's mass, the bremsstrahlung process, or, as the astronomers call it, the free-free transition, is the dominant mechanism determining the opacity of the star. In this particular case the free-free transition is rather strongly temperature-dependent and contributes an additional third or fourth power to the dependence of the luminosity. For a star of roughly the sun's mass the luminosity goes as

$$L \propto G^{7-8} \qquad \text{(for a star of solar mass)} \tag{17}$$

The simplification of holding the radius constant is in no sense justified. The central temperature of a star is determined by the temperature at which nuclear reactions go. This temperature does not depend on G. What really happens is that the radius rather than the central temperature changes. However, it turns out from more careful considerations that the luminosity dependence on the gravitational constant we have obtained is approximately correct whether the radius or the central temperature changes.

Thus for a star of the sun's mass the luminosity is a rather sensitive function of G . The change of luminosity is of the order of eight times the change in the gravitational constant. If the G variation is of the order of 3 parts in 10^{11} per year, then the luminosity variation will be given by

$$\frac{\delta L}{L} \simeq 8 \frac{\delta G}{G} \sim 2.5 \times 10^{-10}/\text{year} \tag{18}$$

In a period of four billion years this would make this change in luminosity the order of 100%.

A G variation of this rate could play a rather important role in stellar evolution rates and lead to a serious discrepancy in the presently determined stellar evolutionary age of stars.

It is another more difficult matter to determine how the radius changes. It is somewhat uncertain as to how the radius would vary.

Now I will consider the history of the galaxy as we see it and discuss how an accelerated stellar evolution in the past would affect our observations. We picture our galaxy as originally a large mass of hydrogen which in a very short time after its formation produced an initial population of stars. These stars are called Population II. They are found in globular clusters and as field stars of high velocity. The reason for believing that the formation of Population II stars all happened in a rather short time is that these high random velocities appear to reflect the initial turbulent motions of the gas.

If this is the case it might mean that Population II was formed in a time which is of the order of the characteristic time for turbulence to damp out in the galaxy. This time was probably under 200 million years.

Another possibility is that the initial population was so bright and active with OB type and other massive stars that the turbulence was driven by the radiation from these stars. Then it might have been maintained for as long as one billion years. The turbulence could not have been maintained by bright stars of the initial population for much longer than one billion years, for massive stars do not live long. The stars that have the right ultraviolet spectrum to drive turbulence have a short lifetime. Something up to one billion years for the time of formation of the principal part of Population II seems to be indicated.

This time scale seems to fit reasonably well with Salpeter's suggestion that the rate of formation of stars is proportional to the amount of gas present. This would lead to something like two-tenths of the total life of the galaxy for the halo formation period. This is somewhat longer than one billion years. It is of the order of two billion years.

Another thing which characterizes Population II is the very small amount of heavy elements in these stars. They seem to have been formed out of hydrogen.

On the other hand the thing that characterizes Population I stars is that they all seem to have about the full solar heavy element abundance. The measurements which Arp and others have made indicate that the fractional abundance of heavy elements increases in time roughly as given in Figure 3. The logarithm of the ratio of metal abundance to hydrogen is plotted on some arbitrary time scale. The metal abundance rises very rapidly to solar abundance. The very old cluster NGC - 188 shown on the diagram after the formation of Population II seems to have as much heavy element content as the sun.

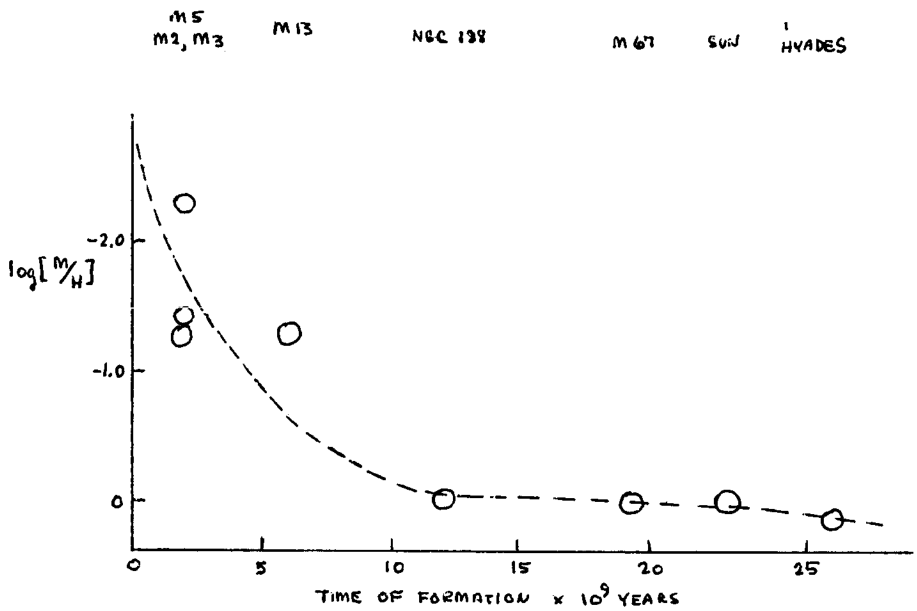


FIGURE 3. The metal-to-hydrogen ratio, relative to that in the sun as a function of the time at which the star clusters formed (Arp).

There is considerable uncertainty about the size of the numbers in Figure 3 because it is quite difficult to measure small changes in heavy element abundance. But the indication is that the principal heavy element formation occurred in stars in the halo population. This is the reason that Population II and halo stars are associated. After the halo population was completed the lower velocity stars started appearing. They seem to have most of the heavy elements in them from the beginning. Very little more has to be added later on.

In the initial population, there were stars of a type no longer found in the galaxy, namely massive Population II stars. We no longer see them and do not know their properties on the basis of observations. They may have been supernova prone; they may have been the primary source of the heavy elements. With the assumption of a larger gravitational interaction in the past, this could have affected the stability of stars in such a way as to make supernova formation likely.

One has a way of dating stars which essentially depends on asking how long it takes them to burn their hydrogen. They start out with hydrogen and some heavy elements and the nuclear reactions occur at the core. The burning in the core keeps moving out as the hydrogen at the center is used up. As it moves out, the luminosity increases slightly. Finally these stars reach a phase where they start changing their form completely. A great expansion takes place and they turn into red giants.

Figure 4 is the familiar Hertzsprung-Russell diagram for stars. The logarithm of the luminosity is plotted against the logarithm of the surface temperature. Luminosity is a measure of how rapidly a star is radiating. The sun's luminosity is 1 in these units. The sun's temperature is indicated in the diagram.

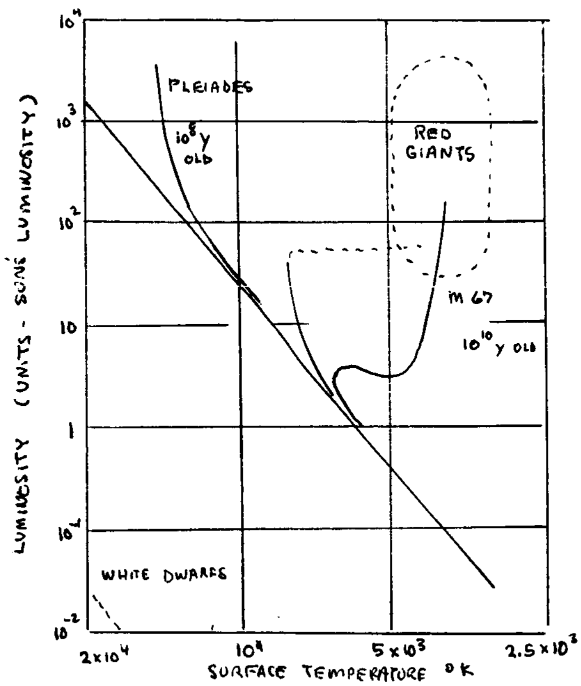


FIGURE 4. Hertzsprung-Russell diagram of stellar distributions.

When a star is first formed it falls on the straight line referred to as the main sequence. The mass of the star determines its initial position on the main sequence. The more massive stars begin with higher luminosity and temperature. After the star has burned a certain faction, some say 20 to 30 percent of its hydrogen, it moves off the main sequence rather rapidly over into the red giant region. After this, it is not completely clear what happens. Presumably the star ends up eventually as a white dwarf. Or else, as mass loss takes place, it might do something quite complicated.

The procedure, then, for determining the age of a star in terms of its luminosity involves the time it takes for a star of a particular mass to burn some 20 to 30 percent of its hydrogen. This percentage for the star to burn before moving off the main sequence is calculated from stellar evolution theory. A set of stars all made at the same time, the Pleiades cluster for example, would initially have fallen along the main sequence. But the massive luminous stars burn up their hydrogen rapidly and move off the main sequence into the red giant region. So we find that at the present time the Pleiades do not fall completely on the main sequence, but fall along the curve shown. The shape of this curve (the point at which it leaves the main sequence) enables us to determine the age for this cluster of stars. This determination should be as good as our 20 to 30 percent figure.

The Pleiades are quite young. In the case of an older cluster, the massive stars are already dead. They have turned into red giants. Then, presumably they evolved rather quickly ending up as white dwarfs. The stars of a very old cluster, M 67, are given in the lowest curve.

These curves do not represent a time sequence. A particular star does not move along the curve. The curves represent the distribution of stars at some particular time. This distribution enables one to obtain a measure of the age of the cluster. This is the basis for stellar evolutionary ages. If gravitation is changing with time, the ages determined this way will be faulty.

Figure 5 is a table of ages of various groups in the galaxy. The globular clusters are among the oldest stars that we know. Their stellar evolutionary ages have been given as 25 or 26 billion years in three cases at least, and 22 I think in a fourth case. These are the figures I will use.

We will take 25 billion years to be the age determined assuming a constant G . Then with the accelerated evolution that would result from increased gravitation in the past this comes down to about 7.8 billion years. This is still a little uncomfortably close to the age of the universe that one obtains from the Hubble expansion. The Hubble expansion age for the universe, assuming an evolving universe of the closed or flat type, is about 8 billion years.

Let me run through the effect of a varying G on some

Object	Type of age	General relativity (Constant G)	(Positive curvature) Brans-Dicke $\omega=6$
Globular cluster	Stellar evolution	25	7.8
Old galactic cluster NGC 188	Stellar evolution	16	7.0
Sun	Stellar evolution	4-15	2.5-6.9
Sun	Radioactivity	4.5	4.5
Galactic system	From depletion of hydrogen gas	5-12	5-12
Elliptical galaxies	Stellar evolution (mean age)	10-16	5.5-7.0
Uranium 25% prompt	Time of first formation	11.1	11.1
"50%"	Time of first formation	7.5	7.5
Universe	Hubble (galactic expansion)	13.0	15.0
Universe (flat)	Based on Hubble age	8.6	...
Universe (closed)	Based on Hubble age	$\lesssim 8.6$	8.0
Universe (open)	Based on Hubble age	< 13 > 8.6	...

FIGURE 5. Age (in unit 10^9 years)

of the other ages as shown in Figure 5. In the old clusters I mentioned before, 16 billion years goes to 7. Notice the very tight compression to only a 0.8 billion year difference between the globular cluster and the old galactic cluster, NGC 188, which differ very greatly in their composition.

There is not a good evolutionary age for the sun because of the fact we do not know its helium abundance. It is believed that the sun's evolutionary age could lie anywhere in the range shown. The radioactivity age for the sun, however, might be taken to be the age of the meteorites. This is of the order of 4.5 billion years.

The age for our galactic system is based on Salpeter's ideas about the way the galaxy evolves. His assumption is that stars are formed at a rate proportional to the amount of hydrogen present. From the present e-folding rate for the condensation of hydrogen into stars we obtain about 5 to 12 billion years for the age of the galaxy. In the case of elliptical galaxies there is a determination of an age based on some work of Hoyle and Crampin (5). From the color distribution of the elliptical galaxies, one determines an evolutionary age of the order of 16 billion years. This age becomes 5.5 to 7 billion years

---

(5) J. Crampin and F. Hoyle, Monthly Notices Royal Astronomical Society, 122, 27 (1961).

in the revised time scale for stronger gravity in the past.

Well, then, one of the effects of a stronger gravitational constant in the past is the shifting of these old stars down to younger ages. This gets them under the age of the galaxy based upon an expanding model of the universe. But as I pointed out, there is considerable uncertainty about these ages. I remember it was only some 4 or 5 years ago that the globular clusters were said to be some 6.5 billion years old. So you see that this whole thing is considerably in flux and one cannot be too much impressed by these numbers.

The age given by the decay of uranium is determined from the relative abundance of uranium, that is, the uranium per gram of hydrogen in the interstellar medium as measured in meteorites or in samples of the earth's crust. The assumption is made that uranium production is such that its abundance is increasing linearly with time, which is the sort of thing which goes with Salpeter's model. In addition there was some prompt production of uranium in connection with the halo population for reasons which I mentioned before. In one case I assume 25 percent prompt; in the other 50 percent prompt. More recently I have calculated an age with the assumption that the uranium is produced by the halo population with distribution curves of the kind shown in Figure 6. These three distribution curves give 9, 7.7 and 7.2 billion years for the first origin of uranium.

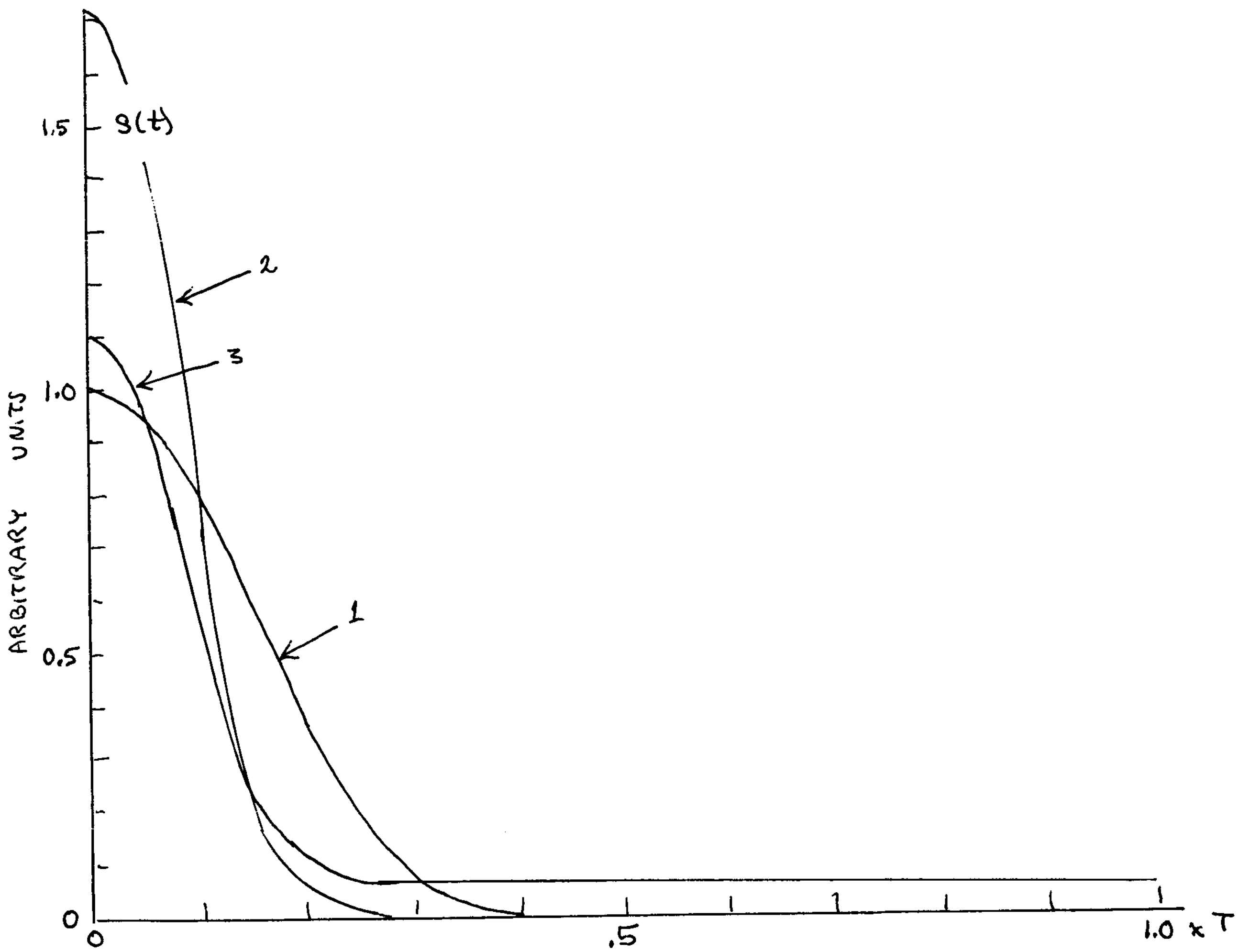


FIGURE 6. Uranium production rate (arbitrary units). T represents the present age of the galaxy.

I have avoided using the ratio of uranium to thorium abundances since they have different chemistries and long half lives. I do not believe anything can be concluded from thorium abundances. I have used only U-235 and U-238 abundances.

Since U-235 has a relatively short lifetime, this age determination is very insensitive to what we assume about the initial abundances. The initial formation ratio of U-235 to U-238 does not matter. Furthermore, if the uranium is made with an R process, then the formation ratio of U-235 to U-238 ought to be a fixed ratio independent of the local conditions in the region in which the sun was formed in the galaxy.

It is possible that these time distribution curves for the formation of uranium are incorrect. For example, if shortly before the solar system formed a supernova occurred nearby, then some of the uranium in the sun could have been produced at that particular time. This could have biased the U-235, U-238 ratio. Because of the short half-life of U-235, a significant portion of it, found in the solar system, might have been formed that way. On the other hand U-238 has a longer lifetime and may have accumulated to a much greater extent from the past. Therefore its abundance is less sensitive to recent events in the solar neighborhood. In this way the details of the formation curve can play an important role in determining the age that we get. It should be noted though that at the time of formation of the solar system the U^{235}/U^{238} ratio was 0.34. Thus a very significant faction of uranium must be assumed to have been formed just prior to this solar system if one is to conclude that this is the explanation for this short time scale.

The uranium ages I have given differ quite a bit from those of Hoyle and Fowler. They make a quite different assumption about the distribution in time of the formation of uranium. Hoyle and Fowler assume that uranium is produced in a kind of supernova which cannot occur until some 4 or 5 billion years after initial stellar condensation. So there is about a 4 or 5 billion year waiting time. They also assumed that the fractional composition of uranium relative to hydrogen varies at a rate proportional to the amount of hydrogen gas present and, as a consequence, that the uranium abundance is an increasing function of time as shown in curve (a) Figure 7. That was a mistake. They should have taken the uranium composition relative to the total amount of hydrogen initially present. That gives the linearly increasing rate in curve (b) Figure 7. However, the reason for the discrepancy is the 4-5 billion year delay introduced initially.

My ages also differ from those of Cameron. In Cameron's model the primary (light) elements are assumed to be formed at the same rate as star formation which is taken to be a decreasing exponential. The secondary elements are formed at a rate dependent on the build up of the primary elements.

For uranium, the production rate is complicated and rather different in terms of history from that of Fowler and Hoyle. But the conclusions about the age of the elements are essentially the same as theirs.

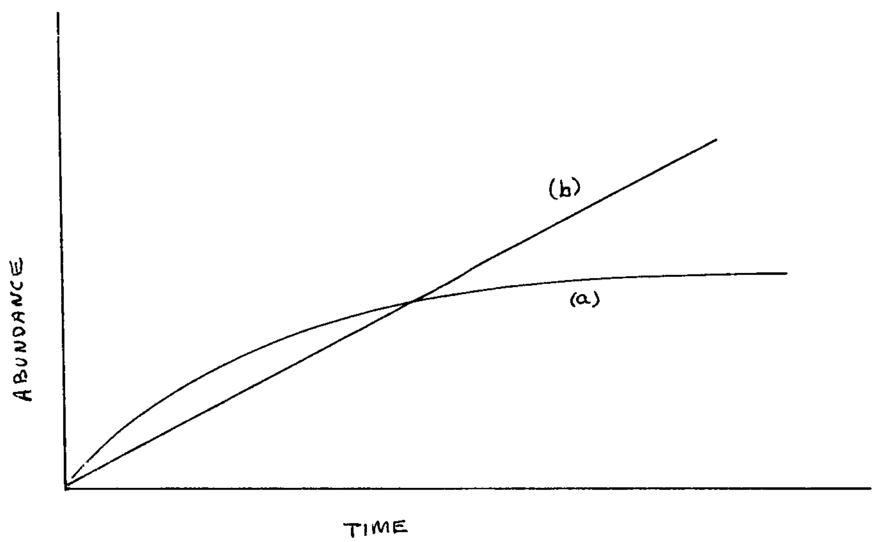


FIGURE 7. Uranium abundance as a function of time.

Now what about the problem of the higher temperatures in the past? I find that if we assume that the black body radiation characteristics of the earth have not changed, a certain change in the solar temperature would lead to a corresponding change in the earth's temperature. If we go back some 4 to 5 billion years, the temperature would rise from about 300° K up to about the boiling point of water.

This is not corrected for solar evolution and the motion of the sun along the main sequence. This would tend to pull the temperature down but not by a great deal. It might come down some 10° K, so that it would be 80° K or 90° K, still approaching the boiling point of water.

The total factor which the luminosity changes due to changing G is about two. This is assuming a linear extrapolation. We take

$$\delta G/G = 3 \times 10^{-11} \text{ /year.} \tag{5}$$

The luminosity changes by a factor 7-8 times as great. So that in 4 billion years the luminosity change would be

$$\delta L/L = (4 \times 10^{9})\ (3 \times 10^{-11})\ (\ 7\ ) \sim 1 \tag{6}$$

The possibility of the earth's temperature approaching the boiling point of water 4 billion years ago is contrary to the argument of Urey that the temperature of the earth has been never more than 300° K. This argument is based on the presence of certain volatile elements still in the earth's crust. I am not terribly convinced by this argument. Only a small part of the earth at the surface has been exposed to erosion. In any case, suppose the volatiles did come out. Where would they go? Only back into the crust eventually.

An important influence on the effect of a change in the sun's luminosity on the temperature of the earth is the water vapor in the atmosphere. The effects of increased water vapor work in two directions at once. One is the increased greenhouse effect leading to a rise in the temperature. The other is increased albedo and better heat transfer to high latitudes, leading to a decrease in the temperature. I think I would argue this way. With increased radiation the first thing I would expect would be that the surface temperature would go up somewhat, leading to a higher vapor pressure, and an increased greenhouse effect, but, on the other hand, also increased cloudiness.

However, there is an argument that there must be very large amounts of water vapor in the atmosphere before the cloud pattern changes very much. I will make the argument, but I am not sure it is right. It is that in the convection of the atmosphere there are both upgoing air and downgoing air currents. These occupy roughly equal areas, so that one would expect roughly 50 percent cloud cover over a wide range of water vapor content.

It is possible that with an increase of radiation from the sun, the difference in the radiation absorbed at the equator and the pole would increase. This could result in more circulation and increased cloud cover where it is most effective. If the atmosphere approached something like 80 or 90 percent water vapor, then the circulation pattern would change in a rather interesting way. There would no longer be the convection cells of the kind that we are familiar with. The water vapor would rise in the equatorial regions and fall as rain in the polar regions. Then there could be nearly 100 percent cloud cover. This is how the ice ages were once explained, by increased cloudiness and by increased rainfall.

These possible changes in the surface temperature of the earth might be significant for biological considerations. If the earth was too hot in the past, living organisms would have been uncomfortable. This is the only real sensitive test of past temperatures that I can think of.

But in the absence of further evidence, I think the moral is that the atmosphere is complicated. One can not make any very firm predictions concerning the effect of an increased luminosity of the sun in the past on surface temperature. We cannot be sure how much the surface temperature would have changed.

In the case of the moon things are certainly much more clear. We can predict unambiguously a higher surface temperature for the moon in the past, approaching some 70 to 100 $^\circ$C some 4½ billion years ago.

Another interesting geophysical effect to be expected, associated with a decreasing gravitational "constant," is a steadily expanding earth. The earth is substantially compressed by the gravitational force. As this force gets weaker with time the earth expands. With a rate of decrease of the gravitational constant of 3×10^{-11} per year, the earth would be expected to expand in circumference by approximately 150 kilometers per billion years. This expansion rate is based on what one knows about the present amount of the compression of the earth. The corresponding number for the moon is of the order of 1 kilometer change in the moon's circumference, per billion years.

What is the evidence on the expansion of the earth? The traditional explanation for mountain formation, the classical one, is a contracting earth with the crust buckling and producing mountains. This classic explanation for mountain formation has fallen somewhat into disfavor in recent years. Many geophysicists no longer take this explanation seriously.

There are some striking indications of something like an expansion in the earth, but again the problem is one of the ambiguity of the evidence. Figure 8 shows an old classic problem faced by the geologists. This is a picture due to Carey, who has been able to explain many geological features as resulting from a large expansion of the earth. According to Carey, as the earth expanded, a great big crack opened up along what is now the

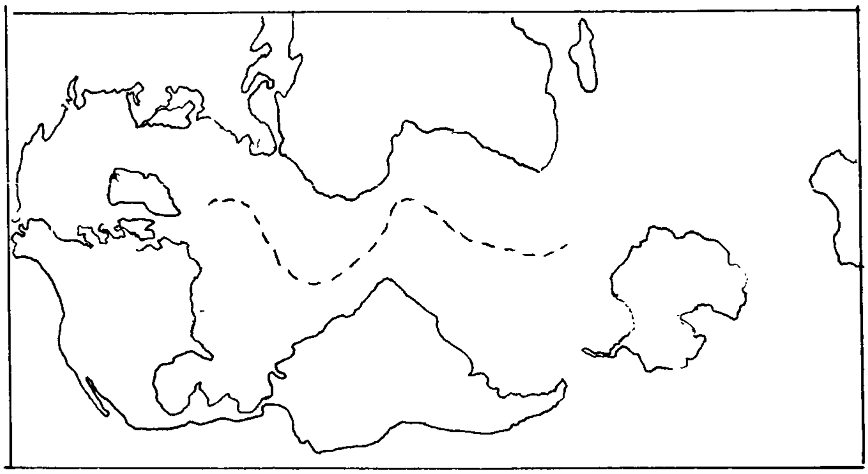


FIGURE 8. Map showing the coast lines of the Americas and Europe-Africa in relation to the Mid-Atlantic ridge (from W. S. Carey).

western coast of Africa and the continents of Africa and South America pulled apart. Figure 9 shows the rather good fit between the boundaries of these two continents. The outline shown is the continental margin, that is, the continental shelf boundary.

Among other geologists who have adopted variations of this explanation are W. Egyed, T. S. Wilson, and Bruce C. Heezen. Wilson and Heezen have suggested that earth expansion may be an explanation for the global system of rift valleys, such as the medial crack system along the Mid-Atlantic ridge.

Unfortunately, an expanding earth is not the only possible explanation for the geological features pointed out by Carey. The old explanation of A. Wegener (1915) and A. L. DuToit (1937) involving continental drift is a possibility. This idea was placed on a more reasonable basis when it was recognized that a convective mantle could result in motion of continental masses. In addition to the geological features mentioned above, there is other evidence for the relative motions of continents. Recent paleomagnetic data, much of it taken and studied by S. K. Runcorn, has given evidence for continental drift (which would not require an expanding earth but probably would require a convective mantle).

One might argue that the coincidence between coast lines does not mean anything. With all the many complicated coasts one might always find some coast lines which would fit together. However, a compelling argument for a common origin of these two coastlines is the existence of the mid-Atlantic ridge 1 to 2 kilometers high shown in Figure 8. This mid-Atlantic ridge is quite accurately halfway in between the continental coasts.

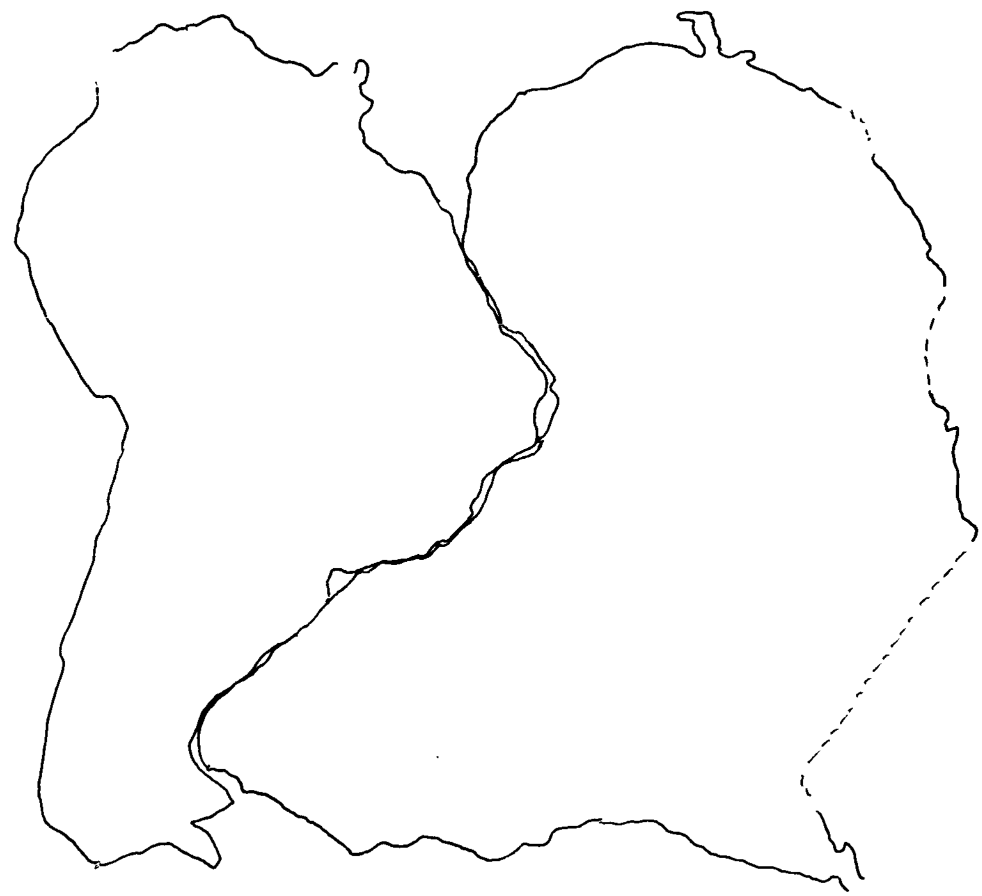


FIGURE 9. Map showing the fit between South America and Africa at the 2000 mile isobath along one slope below the edge of the continental shield (from W. S. Carey).

It rather accurately represents a medial ridge down the Atlantic Ocean basin.

As mentioned above, this ridge has along much of its length a medial crack which is quite large, some kilometers wide.

This is as if the earth were really pulling apart there, thus forming the crack. It is not a continuous crack along the whole length of the ridge, but it does seem to occupy a large part of the total length.

This feature of a mid-ocean ridge is not limited to the Atlantic. It exists in all the oceans. It is shown in Figure 8 continuing around Africa. It goes into the Indian Ocean, and extends around into the Pacific. In fact, there is a connection with the Gulf of California. In the Gulf of California there is a crack running north and south which has gotten into the land.

Another interesting thing is that the Atlantic ridge seems to run right through Greenland, The crack is on dry land there. You can walk around and look at it. The land is up quite high. All the igneous activities that one sees in Greenland seem to be associated with the fact that it is part of the mid-Atlantic ridge and that this crack runs through it.

This is a rather compelling argument, I think, that these continents were at some time closer together and that they were associated in some way. The explanations for this are where the disagreements come. Carey suggests that the whole earth has expanded an enormous amount, much more than I would have liked to have with gravity getting weaker. We cannot get very much expansion from our small rate of decrease of gravity.

It is unlikely that changes in the structure of matter toward the center of the earth would lead to a disproportionately large radius change. The usual assumption is that the inner core is a solid form of iron and nickel, and the outer core is liquid. A change in G could change the phase boundary and cause the region of melting to shift. But I do not expect anything very discontinuous to happen. Even if the inner core were a solid phase of the same composition as the outer core, I do not think there could be an abrupt change in radius. The effect of an expansion is always one of absorbing heat and shoving the reaction which provides the expansion back in the direction to turn it off. It is not something for which an instability develops. This enormous required expansion is one of the very serious problems connected with Carey's ideas.

Another direct observational bit of evidence against Carey's ideas is the fact that if the earth were expanding at the rate at which Carey says, it would lead to some very noticeable effects in the motion of the moon relative to the earth's rotation. The day would be slowing down at a rate decidedly greater than has been observed over historical times.

The explanation based on mantle convection for the apparent drift of continents is largely due to Vening Meinesz. According to Vening Meinesz, the early convective cooling of the earth was with a simple system of convective cells which became more numerous as the core of the earth developed. In this model, the mantle of the earth, although one would think that normally it is solid, is an almost viscous liquid continually convecting. As stated above, the convection has caused continents to pull apart, forming such oceans as the Atlantic. This is assumed to have happened in the recent past. Of the order of 100 to 200 million years ago they were joined together. They have separated since then.

As was mentioned briefly, the idea that the continents are moving around, because they are floating on the mantle, goes back to Wegner. He explained the Ice Ages by having continents drift up to the North Pole where they have an Ice Age and then drift away again. He would have had the continents drifting around like bits of wood in a quiet pond of water.

There is a number of interesting things one can say about the effects of convection in the mantle, if it exists. If there is an uprising cell along the mid-ocean ridge we would expect a higher heat flow. Well, one does see a higher heat flow.Also, if material is rising here we might expect that this ocean bottom is rather recent. There would not be much in the way of sediment on it. Well, there is not much in the way of sediment. This is a rather surprising thing about all the oceans. There is very little in the way of sediments on the bottom. Using present sedimentation rates, one would expect that there would be considerably more than there actually is.

As mentioned above, the recent measurements on paleomagnetism rather strongly indicate continental drift is going on. But if there is continental drift going on associated with mantle convection, I think it would be very difficult to say anything about a general expansion. Effects of a general expansion are too small compared to these much larger effects, and they are easily masked.

The convection itself might be associated with a decreasing gravitational constant. This is because of the fact that as you take the pressure off the earth in the interior, the melting point decreases. As the melting point gets closer and closer to the temperature that exists there is either local melting or at least the viscosity falls to the point where convection starts. So there may be actually a connection between convection and a weakening gravitational constant.

A third explanation which has been given for continental drift is that in the early days of the formation of the earth there was a rather large amount of convection in the interior of the earth. This was either in the form of a solid mantle or else a molten earth associated with the heat of the initial radio-activity and the heat energy associated with the compaction, that is, the gravitational energy. The convection in the original earth produced the large convection cells that determined the land mass distribution. Then these convection cells disappeared. So that the land mass distribution that we have now is a fossil remnant of early convection cells. This is an explanation which is favored in some quarters.

If it is true that this convection is not going on now, then there are quite reasonable explanations for the oceans' system of cracks. They might be due to a general expansion connected with weakening gravity. But if the convection is going on now, I think the direct effects of general expansion in producing such features are minor.

In connection with the moon's expansion, there is a similar situation except that there is no evidence for convection in the case of the moon. If there were convection, faulting of the surface would be expected. One would expect to find craters sliced in two, one half sliding with respect to the other half. These do not seem to appear. It is clear that the moon might actually be a better place for looking for expansion effects even though the expected expansion is much smaller. The effects of expansion would be expected to appear in the form of surface cracks or magma flows. Magma flows might be expected to result from an expansion of the interior, the only part requiring expansion. This could result in the internal low melting point components forcing their way out through cracks in a rigid crust.

(See U.S. Air Force Lunar Atlas: Plate D3-a, Archimedes.)

FIGURE 10. The moon's crust showing characteristic maria and craters

Figure 10 shows the moon's surface with its characteristic maria. We are all familiar with these large dark areas on the moon which could be lava flows. It has been suggested by T.Gold that these are seas of dust. I doubt the validity of the "seas of dust" explanation. There are a number of craters flooded inside and out to the same level, as nearly as one can tell from the shadow measurements of height. It is very difficult for me to conceive how dust would establish hydrostatic equilibrium, filling up inside the crater to the correct height. This seems to suggest more directly a fluid, connected through surface fissures to a common sub-surface magma pool.

(See U.S. Air Force Lunar Atlas: Plate C2-b)

FIGURE 11. Photograph of moon's surface showing a gash.

Figure 11 shows another feature which was described at one time as a gash caused by a meteorite fragment. I think that one can soon convince oneself that a large high velocity projectile would not make a gash like this, but would produce an intense shock wave that could result in a crater-like formation. I think a much more reasonable explanation for this particular formation is a crack in the surface, a fissure filled by magma from the interior.

Another effect of gravity getting weaker with time that would be expected is a gradual slowing of the moon in its orbit about the earth. This should lead to a discrepancy in the lunar position computed on the basis of constant G. Figure 12 shows

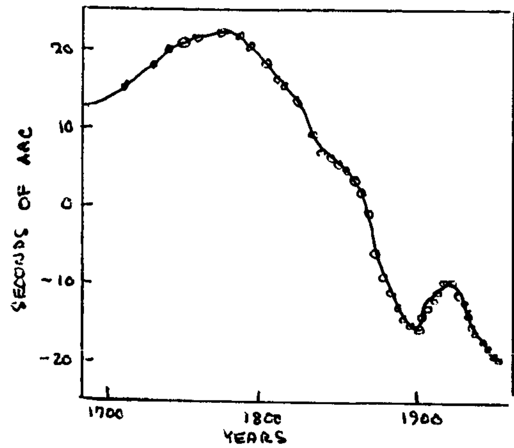


FIGURE 12.The discrepancy f (T) of the moon's longitude based on occultations

the lunar discrepancy curve as observed from telescope observations for the last 200 years. The error in the moon's position, that is, the observed longitude minus computed longitude, is given as a function of time as determined by the earth's rotation rate. There are at least two effects contributing to this error. One is an irregular fluctuation effect, usually assumed to be due to an irregularity in the earth's rotation rate.(See Lecture 12 for an alternative explanation.) The other is a quadratic effect indicated by the parabolic shape of this curve. The quadratic effect is associated, in part at least, with the tidal interactions between the earth and the moon which slow the earth's rotation down and cause the moon to move out to a bigger radius with a longer period. One can eliminate the fluctuation effect due to the earth's erratic rotation and simply look at the tidal slowing down of the moon by combining the moon's observations with observations of the sun's and Mercury's positions.

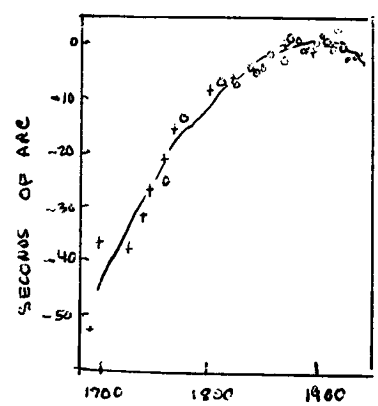


FIGURE 13. Weighted discrepancy differences: o for the sun, + for Mercury (prior to 1740 these depend on the extrapolated longitude of the sun). (From MacDonald and Munk)

Figure 13 shows how this combined data looks. The curve is a parabolic arc without fluctuations, the irregularities in the earth's rotation having been taken out. From this curve one can determine the rate at which the moon has been slowing as a result of the effect of tidal interactions only. The tidal effects can be computed directly from the observations.

This slowing of the moon's motion, due to a tidal interaction with the earth, implies also a slowing of the earth's rotation. This is not the only tidal interaction with the earth's rotation which requires notice. There are other tidal effects one needs besides this. Two other tidal interactions affect the earth's rotation rate. In addition to the tidal coupling of the moon with the earth, the tides raised on the earth by the sun affect the earth's rotation. Also there is the atmospheric tidal couple. As mentioned above, from telescopic observations one can obtain the tidal slowing of the moon's motion and the resulting slowing of the earth's rotation rate. Assuming that the tidal slowing of the earth's rotation is proportional to the tidal driving force, the tidal slowing of the earth due to the sun can be computed. Also the measured atmospheric pressure fluctuations allow the atmospheric tidal speeding of the earth's rotation rate to be computed. Combining all these effects, we can compute the expected slowing down of the earth's rotation rate from all the tidal interactions.

There is another expected effect on the earth's rotation rate connected with the fluctuation of sea level. Figure 14, taken from Fairbridge (6), shows the kind of data that one has

---

(6) R. W. Fairbridge, Physics and Chemistry of the Earth, Vol. 4, Ahrens, Press, etal ed., p. 158.

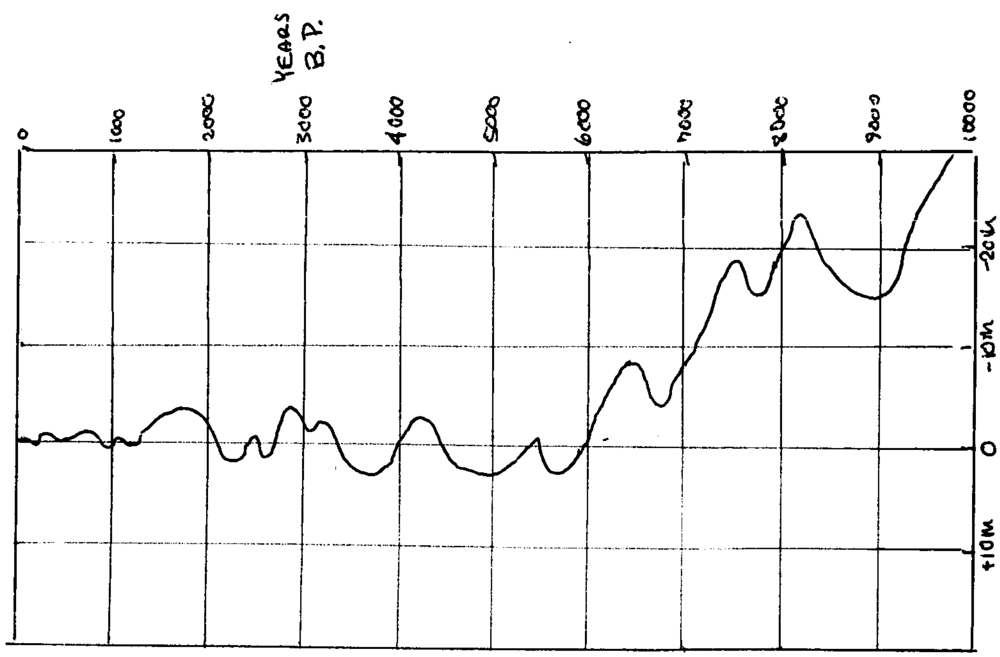


FIGURE 14. Historical variation in sea level (from R. W. Fairbridge).

on sea level fluctuation which would have affected the moment of inertia of the earth. The sea level in the past is determined by radiocarbon dating of coastal shells. The old eclipse observations on the earth's rotation rate are primarily in the period when the sea level was most rapidly changing. So we should take into account the effects of the variations of sea level occurring over that period of time.

If sea level should rise by 1 cm as a result of arctic or anarctic ice melting, the earth would be expected to rotate more slowly by a part in 10^9, after including the effect of elastic distortion of the earth, but assuming no isostatic adjustment of the crust. It is probably more reasonable to assume a substantial isostatic compensation of the shape of the earth, and to assume that the effect on the earth's rotation rate is proportional to the change in sea level, proportional with a proportionality constant to be determined.

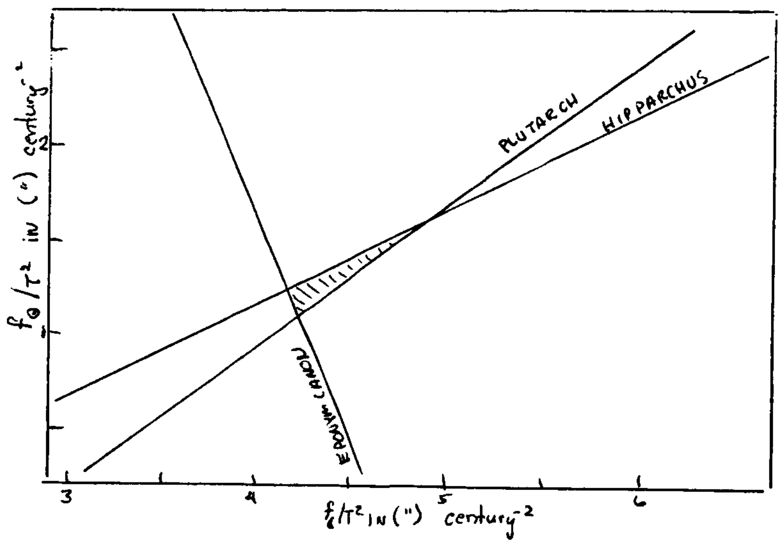


FIGURE 15. Fotheringham's (1920) summary of consistent values of $f_{☾}/T^2$ and $f_{\odot}/T^2$ for various ancient solar eclipses.

Figure 15 shows the kind of data that one has to use in order to obtain an observational value for the secular acceleration of the earth's rotation and the moon's motion. The data must be taken over long periods of time because of the irregularities in the earth's rotation rate. This means going back to the classical eclipse observation of the Babylonians and the Greeks. From any one eclipse observation, such as the one described by Hipparchus, one obtains a linear relation between the secular accleration of the sun and that of the moon. The secular acceleration of the sun is a measure of the acceleration of the earth's rotation rate. (Two lines, representing upper and lower limits, are shown.) These are the curves shown in Figure 15. While the secular acceleration of the sun is a measure of the earth's rotational acceleration, the moon's acceleration is due to both effects, and acceleration of of the moon and a slowing of the clock given by the earth's rotation.

In the framework of this analysis, a gradual slowing of the moon's motion and planetary motion, due to a gradual weakening of the gravitational interaction, would appear as an unexplained residual speeding of the earth's rotation rate, after making allowances for all known effects on the earth's rotation. By combining modern telescope observations and the information obtained from these eclipse observations, one can get a secular acceleration for the sun and hence an effective average acceleration of the earth's rotation. Subtracting from this the three tidal accelerations determined from the telescope observation and barometric pressure fluctuation, yields a residual discrepancy which is just about the right size to be a consequence of gravity getting weaker. It is found that the four best eclipse observations considered by Fotheringham, the only ones really worth discussing, are made more consistent if an allowance for sea level fluctuations is included. The resulting derived proportionality between sea level variation and the earth's rotation rate allows correction for sea level fluctuations to be included. After including this correction the earth's rotation rate exhibits an even larger residual acceleration.

Unfortunately, there is one thing that we can't really be sure about, and that is what the earth's core has been doing. The earth's core could have coupled with the mantle over a very long period of time, transferring momentum from the core to the rest of the earth. One could account for the discrepancy this way. This possibility seems to be the chief unknown that exists in the data and theory.

I would like to mention one more place where it seems to me there might be a quite interesting effect. This is an effect connected with Jupiter's interior.

There may be a real problem of accounting for the magnetic field of Jupiter. The relaxation time for an electrical current in the interior of Jupiter to die out is sufficiently short that Jupiter should not have a primordial magnetic field left over. One must account for Jupiter's magnetic field in terms of a mechanism presently stirring up the interior in a magneto-hydrodynamic way generating this magnetic field. The energy required to do this appears to be quite large. This is because the deep interior of Jupiter is, as far as one can tell, hydrogen in a degenerate form. It has a very good thermal conductivity. Therefore, it is quite difficult to drive mass convective currents with heat flow. While convection in the outer part may be possible, this part is probably not a good electrical conductor.

It is possible that there might be a non-degenerate solid core of heavy elements a few times the earth's mass. This is the suggestion of DeMarcus. Even with all the radioactivity that might be in such a core it is quite difficult to get a large enough heat flow to produce temperature gradients sufficient for convection in the outer part of the metallic hydrogen phase.

It is conceivable that the field is produced in an iron core in the inner core. But this does not fit what we know about Jupiter's magnetic field. The radio measurements suggest that the field is nothing like a centered dipole magnetic field. It appears to be way off to one side, and rather localized. Hence it is not likely that the magnetic field is produced at the center.

A varying gravitational constant provides a possible mechanism for driving currents outside the core. This depends on the existence of a phase change going from degenerate to non-degenerate hydrogen at some distance from the center, 0.6 or 0.7 of the radius of Jupiter. If G. changes, the radius at which the phase change occurs should change as well. The radius of the phase discontinuity and the resulting discontinuity in density should move steadily inward as gravitation weakens.

This density change leads to a difference in the rotation rate of the inner part of Jupiter relative to the outer part. The moment of inertia of the material involved in the phase change does not scale properly to maintain rigid body rotation. Conservation of momentum leads to the inner part rotating more rapidly than the outer part.

How much energy is available from this? From the G variation rate we have been discussing I estimate that it might be 20 to 100 times as much energy as is available from radioactivity. This energy would be made available by some mechanism of damping out the differential rotation through magnetic coupling between the two conducting shells. Furthermore, this mechanism provides a shearing motion for the production of a magnetic field.

Magnetic field lines cutting through the radius of phase change would be stretched out, wound like yarn on a ball, until magnetic pressure and tension effects become important. The resulting magnetic forces could induce turbulence in the boundary region.

It is quite conceivable that if anything like this should occur, it could be an important mechanism for stirring up the interior of Jupiter by producing a differential rotation of the interior.

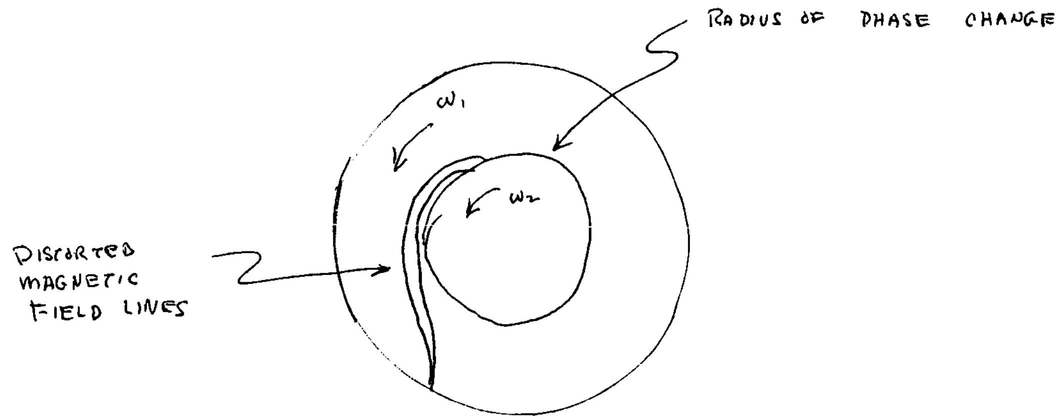


FIGURE 16. Schematic diagram of possible magnetic field generation in Jupiter resulting from differential rotation.

I cannot give a detailed mechanism for the production of a magnetic field in this manner. It is difficult to understand how a magnetic field is sustained in any celestial body. But I could conceive of a situation where this shear motion could be coupled to a non-axially symmetric magnetic field. This might involve convective eddies and turbulent eddies, in the shear region (Figure 16).

It is interesting and may be significant that Jupiter does exhibit various rotation rates. The observable features seem to rotate at various rates depending upon latitudes. Also the magnetic field has its own characteristic well-defined rotation rate, about the same as high latitude visual features.

The effects I have discussed in this lecture do not demonstrate the existence of a time varying gravitational constant, but I have not yet been able to find an effect for which the observations rule out such a variation. Rather, in many cases, a varying G provides a possible explanation for a little understood feature of astronomy or geophysics.

Preface

14306

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R. H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution of space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a certain amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and coorect the final manuscript. Hence, we accept responsibility for errors and omissions.

H. Y. Chiu

W. F. Hoffmann

N68-14306

RELATIVITY PRINCIPLES AND THE ROLE OF COORDINATES IN PHYSICS

It is generally agreed that two principles play a role in the formulation of the general theory of relativity. One of these, the principle of equivalence, is usually accepted without question The other, the principle of general relativity, or, as it is sometimes called, the principle of general covariance, on the other hand has served as a topic of heated discussion ever since it was first put forward by Einstein in 1915. Kretchmann[1] was the first to raise objections to the principle and more recently the question has again been discussed quite forcibly by Fock in his recent book.[2]

The principle of general relativity is essentially the statement of certain invariance properties of a class of physical theories and is therefore of interest to the theoretical physicist for several reasons. Perhaps most important of all, an invariance principle associated with a group of transformations usually implies a limitation on the possible types of theories that one can formulate which satisfy it. Therefore it will be of interest to examine the principle and to find limitations which it imposes on the possible forms for the equations of general relativity. In 1918, Noether[3] showed that there is a very close relation between the invariance properties of a given theory and the conservation laws in the theory. For a theory invariant with respect to a group of transformations whose elements are determined by a finite number of parameters, that is, a Lie group, this relation is quite simple; for every parameter there is an associated object which is conserved. If the invariance is with respect to a group whose elements are determined by a number of arbitrary space-time functions such as the gauge group of electrodynamics or, as we shall presently see, the coordinate group of general relativity, the relationship is not so clear. As a consequence, there has been a great deal of discussion in the past few years over the role of the conservation laws associated with general relativity, namely, the conservation of energy and momentum. Therefore a better understanding of the invariance properties of the theory may lead us to a better understanding of the meaning of the associated conservation laws. Finally, as we shall see, the principle of General Relativity is intimately related to the coordinatization of the underlying space-time manifold and hence may shed some light on the role which coordinates and the process of coordinatizing a manifold play in physics.

Before we begin our discussion of the principle of general relativity it will prove helpful to discuss briefly a more familiar principle, that of special relativity. Even here things are not as straightforward as some books would make us believe. The basic intuitive idea is simple enough; we should not be able to distinguish between the totality of frames of reference moving uniformly and rectilinearly with respect to each other by any physical means. The difficulty with this definition lies in the apparently innocent phrase "by any physical means". Almost any physical system, such as a box of gas or an electron, does in fact allow us to single out a particular reference frame among the totality of all those moving uniformly with respect to each other, namely the frame in which it is at rest. Let us try to be more explicit about the term "physical means".

Imagine two identically constructed systems, such as two electrons, two identical boxes of gas, etc., moving uniformly with respect to each other. Call them systems A and B respectively We now imagine two observers or frames of reference, A and B, such that the system A is at rest in frame A and system B is at rest in frame B. Then the principle of special relativity requires that, if the initial state of system <u>as seen by observer A</u> is the same as the state of system B <u>as seen by observer B</u>, then the final state of system A, as seen by observer A, will be the same as the final state of system B, as seen by observer B. Note that observer A never looks at system B and vice versa; they merely compare results of what they see their own systems doing. If observer A were able to make measurements on system B he would, in general, obtain entirely different results from those obtained by observer B from system B. It is in the sense used above that we are to interpret the term "physical means" in our first formulation of the special relativity principle. We can then rephrase our original statement so as to read: there exists no physical system or state of that system which will behave in different ways when placed in one or another of a collection of frames of reference moving uniformly with respect to each other.

There is another formulation of the principle which, at first sight, appears to be fundamentally different from the first. It asserts that the laws of physics can be put into a form which remains unchanged when the various quantities appearing therein are subjected to a Lorentz transformation. In order to be able to refer to it readily we shall call this the principle of Special Covariance. In the first formulation there is no mention of how the various physical quantities transform under Lorentz transformations; we did not even mention Lorentz transformations. In fact, it is not always clear that it is meaningful to talk about the Lorentz transformation of a particular physical quantity.

In order to understand the distinction between the two formulations let us consider two different physical systems and the laws associated with them. One of these systems will be the electromagnetic field, the other a box of hydrogen gas. The laws associated with the electromagnetic field are of course Maxwell's equations. They apply to all conceivable electromagnetic fields: to the field of an electron at rest with respect to an observer as well as one moving uniformly or even arbitrarily with respect to the observer. Furthermore, the way in which the electromagnetic field is measured is independent of the particular field being measured. If we want to know the electric field at a point, we would hold a small test body there and measure the force on it. Another observer, moving uniformly with respect to the first, would measure the electric field he sees in exactly the same way. It is therefore a meaningful question to ask for the relation between the two measurements of the same field. It is a relation which in principle could be verified by observation. Let us now see how our two formulations of the principle of Special Relativity apply to the electromagnetic field and Maxwell's equations.

For our two systems in the first formulation we could take an electron at rest with respect to A and another at rest with respect to B. All that is required is that the field measured by observer A due to electron A be the same as the field measured by observer B due to electron B. However, since both observers are able to measure any and all electromagnetic fields we could interchange the roles of the electrons. A's system would then be an electron at rest with respect to observer B, while B's system would be an electron at rest with respect to observer A. And again, the field which A measures should be the same as that which B measures. A similar duality holds for any system of charges and fields. Since a physical law is a statement about the behavior of a collection of physical systems, all of which possess some common properties (in fact, the physical law is just a statement of these common properties) and since all of the electromagnetic systems which A observes are identical with the totality of electromagnetic systems which B observes it follows that the physical laws governing these systems as formulated by B in terms of the fields he measures in order that the principle of special relativity holds. Furthermore, since it is meaningful to talk about the relation between the values which A would measure for a particular field and those which B would measure for the same field we can talk about the transformation of the laws which A has formulated into the laws which B has formulated. Since the two sets of laws have the same form, the transformation between the two fields must be such as to preserve this form. In this way we are led to the statement of special covariance as we formulated it above. As we well know, it is just the Lorentz transformations which maintain the form of Maxwell's equations and not the Galilean transformations although our first formulation was equally applicable to both types of transformations.

Let me summarize what I have said concerning the electromagnetic field. Observer A looks at the totality of all electromagnetic fields and finds that they satisfy a set of laws which can be written in the form

$$F_{A,\nu}^{\mu\nu} = - j_A^{\mu} \tag{1}$$

and

$$F_{A\mu\nu,\rho} = 0 \tag{2}$$

where $F_A^{\mu\nu}$ is the usual antisymmetric matrix constructed from $\underset{\sim}{E}_A$ and $\underset{\sim}{B}_A$, the electric and magnetic fields respectively measured by A. j_A^{μ} is a column matrix constructed from the current and charge densities measured by A. I call these quantities matrices since at this point I wish to specify only their algebraic and not their transformation properties. Similarly, observer B looks at the totality of all electromagnetic fields and finds that they satisfy a set of laws which, if the principle of Special Relativity is to hold, must perforce have the same form as eqs. (1) and (2) except that the subscript A is replaced by subscript B. Now, since the collection of fields which A measures to verify equations (1) and (2) is the same as the collection of fields which B uses there must exist a relation between $F_A^{\mu\nu}$ and $F_B^{\mu\nu}$ as well as between the spatial and temporal measurements of A and B such that, when we substitute into equations (1) and (2) for the quantities measured by A in terms of the quantities measured by B we obtain the correct equations satisfied by the B quantities. Poincaré[4] and Einstein[5] derived the correct transformation equations relating the A and B quantities. For spatial and temporal measurements they are

$$x_B^{\mu} = \alpha^{\mu}{}_{\nu} x_A^{\nu} + b^{\mu} \tag{3}$$

where the b^{μ} are a set of four numbers and $\alpha^{\mu}{}_{\nu}$ is a matrix satisfying the conditions

$$\alpha^{\mu}{}_{\rho}\alpha^{\nu}{}_{\sigma}\eta_{\mu\nu} = \eta_{\rho\sigma} \tag{4}$$

where $\eta_{\mu\nu}$ is the Minkowski metric in Cartesian coordinates and is given by

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix} \tag{5}$$

The fields are then related by

$$F_B^{\mu\nu} = \alpha^{\mu}{}_{\rho}\alpha^{\nu}{}_{\sigma}F_A^{\rho\sigma} \tag{6}$$

Let us emphasize again that the transformation equations (3) and (6), while derived from the condition that the Maxwell equations (1) and (2) maintain their form when subjected to these transformations, are, in principle at least, subject to direct experimental verification. We see thus that electromagnetic fields satisfy both statements of the principle of special relativity.

Let us now consider the case in which the systems to be examined are boxes of gas. Observer A looks at all possible states of the gas for which he can measure thermodynamic quantities such as temperature, pressure, entropy, etc., and deduces from his measurements that the first and second laws of thermodynamics apply to these states of the gas. Similarly B looks at his boxes of gas and, if the principle of special relativity holds, must find that the thermodynamic quantities he measures must also satisfy the two laws of thermodynamics in exactly the same form as found by A.

There is, however, an essential difference between the electrodynamic and thermodynamic cases. Thermodynamic quantities only have meaning in the rest frame of the system being observed. Thus, any measurement of the temperature of a gas streaming uniformly past the observer or, what is the same thing, for the observer to measure the temperature while holding a thermometer in his hand and running past the gas will, in general, depend on the kind of thermometer employed, how it is orientated with respect to the direction of motion, etc. This is not to say that an observer could not infer from measurements on a moving system what its rest temperature is. The point is that he must interpret these measurements in terms of the rest temperature of the system since this quantity alone depends on thermodynamic state of the system. It is therefore not physically meaningful to talk about the transformation properties of thermodynamic quantities since such transformations could never, even in principle, be verified by observation. Thus the requirement of special covariance, as applied to thermodynamic systems, is without physical content. While it is possible to <u>define</u> transformation laws for thermodynamic quantities so that the laws of thermodynamics retain their form when subjected to these transformations and so thereby formally conform to the requirement of special covariance, such a procedure is without physical content. Formulating thermodynamics in a special covariant form was actually carried out by Planck(6) and Einstein(7) and later elaborated upon by Tolman(8).

While the formulation of thermodynamics in the sense outlined above is without physical content I should point out that the relativistic treatment of an ideal gas is another matter altogether. Here we do not ask about the transformation properties of the various thermodynamic quantities; it is assumed that we always work in the rest frame of the gas. However we ask for the modifications in the equations of state for the gas when the molecules or atoms comprising it are moving at relativistic velocities. Jüttner[9] was the first to work out the case for an ideal, relativistic gas. He proceeded by calculating the partition function, Z, given by

$$Z = e^{-H/kT} dx_1 \ldots dx_{2N} \tag{7}$$

where H is the energy of the system and is a function of the 2N variables $x_1, \ldots, x_{2N}$. This is the usual expression for the partition function as found in all books on statistical mechanics. Now however, instead of taking H to be $\sum_i \frac{1}{2} p_i^{\,2}$ for an ideal gas, Jüttner used the relativistic expression

$$H = \sum_i (m_o^{\,2} c^4 + c^2 \underset{\sim}{p}_i^{\,2})^{\frac{1}{2}} \tag{8}$$

Thermodynamical quantities such as pressure and internal energy are derived in the usual manner by taking appropriate derivatives of (ln Z)/kT. Notice that there is no attempt to modify the usual non-relativistic formulation of statistical mechanics. It may conceivably require modification when we deal, with systems whose components are moving with relativistic velocities but the principle of special relativity offers us no clue as to the nature of the modification. The only thing it suggests is that we replace the non-relativistic expression for H by the relativistic one given in eq. (8).

What morals can we deduce concerning the principle of special relativity from our considerations above? Above all we must conclude that it is not so much a statement about the various physical systems which can exist in the space-time manifold. It says very little about physical systems and the form of the laws which are to describe them. This is especially true for the case of systems which uniquely define a rest frame such as our box of gas did. There our first formulation told us nothing and our second formulation was satisfied in a trivial, non-physical way by merely requiring that the various thermodynamic quantities transform in a manner so as to preserve the form of the laws of thermodynamics. Even for systems exemplified by the electromagnetic field where there are no restrictions on which fields can be measured by which observers, the principle tells us very little unless we add the additional requirement that the laws should be <u>local</u> laws, that is, they should be capable of verification solely by means of measurements made in the immediate neighborhood of a point, and that the transformation laws between the physical quantities appearing in these laws should also be local in the same sense. If we further restrict our systems by requiring that the transformation laws between the quantities appearing therein should be linear and homogeneous, i.e., the quantities should transform as tensors, or tensor densities, spinors, etc., then we very seriously restrict the possible types of systems and physical laws which can occur in nature.

We will conclude this discussion of special relativity with a reformulation in geometrical terms. We note first that, as a consequence of the axiom that the velocity of light is independent of the motion of the source we can conclude that, at every point of the space-time manifold, there is an invariant geometrical object, the light cone. Such an object would exist in a metrical geometry and would consist of the locus of all points in the neighborhood of a given point which are at light-like distances with respect to it. Thus, if x^μ are the coordinates of the point in question, then all other points $(x^\mu + dx^\mu)$ on the light cone originating from x^μ satisfy

$$ds^2 = g_{\mu\nu}(x)\,dx^\mu dx^\nu = 0 \tag{9}$$

where $g_{\mu\nu}(x)$ is the metric at the point x^μ. The principle of special relativity then asserts that the space-time geometry is homogeneous and isotropic. Hence there exists a ten-parameter group of motions which leaves the value of the metric unchanged in the sense described in the chapter on Riemannian geometry. As a consequence we can conclude that the geometry is a flat geometry so that the curvature tensor* satisfies

*See lecture II

$$R_{\mu\nu\rho\sigma} = 0 \tag{10}$$

everywhere. Hence we can always find a coordinate system in which the metric is everywhere equal to the matrix $\eta_{\mu\nu}$, whose components are given by eq. (5). The motions admitted by the geometry have, in this coordinate system, the form given by equation (3). Physical objects are then represented by geometrical objects in this Minkowskian geometry and physical laws are then statements of relations which exist between the various geometrical objects. Of particular interest for physics are the local geometrical objects which have linear, homogeneous transformation laws and the local relations which one can construct between them. We should emphasize that all of our statements are geometrical in nature and hold independent of the particular coordinate system actually employed.

Let me now go on to the General Theory of Relativity. As I mentioned in the beginning of this talk, there are still many disparate views on the subject. Fock has gone so far as to proclaim that "As for the 'general Principle of Relativity,' no such principle exists," and that there is less relativity in "general relativity" than in special relativity. He bases his contention on the fact that, while the flat space-time geometry of Minkowski admits a ten-parameter group of motions whose group is just the Lorentz group, a general Riemannian metric may have no motions at all associated with it. While this is true, it is not at all pertinant since, in the general theory, the metric is no longer taken to be given *a priori* as in the case of special relativity but is to be considered a dynamical quantity along with the other fields of nature. In fact, as we shall see, it is just the requirement of general relativity that forces us to treat the metric in this manner. Fock's objection is then equivalent to asserting that electrodynamics does not satisfy the principle of special relativity because the field of an electron depends upon the state of motion of the observer with respect to the electron. In other words, a particular metric is no more a law of nature in general relativity than is a particular electromagnetic field in special relativity.

As in special relativity, there are really two different formulations of the principle of general relativity. One formulation is analogous to our requirement of special covariance. We shall call it the principle of general covariance. It states that the laws of physics can be put into a form which remains unchanged when the various quantities appearing therein are subjected to an arbitrary coordinate transformation. And, like the principle of special covariance, the principle of general covariance is, by itself, devoid of physical content. Thus it has been argued, as an objection to the principle, that any system of equations which are invariant in the sense defined by special relativity, can be put into what appears to be a generally covariant form by performing a coordinate transformation from a Cartesian coordinate system where $g_{\mu\nu} = \eta_{\mu\nu}$ to an arbitrary coordinate system where now the metric will be some space-time function. Thus we can rewrite our laws by replacing $\eta_{\mu\nu}$ by $g_{\mu\nu}$, ordinary derivatives by covariant derivatives and appending the equations (10) for determining the metric. However, we have introduced the general metric in a rather trivial way, which adds no physical content to the theory. Adding physical content to the theory would require generalizing the $g_{\mu\nu}$ to include non flat metrics for which equation (10) is not satisfied (that is, those not obtainable from the $\eta_{\mu\nu}$ by coordinate transform).

There is another example of this kind of trivial extension of a theory which has a bearing on a proposal of Sakurai[10] to explain the strong couplings of strange particles. We know that the Dirac theory of the electron is invariant with respect to the group of gauge transformations of the first kind:

$$\begin{aligned} \psi^{*\prime} &= e^{-i\alpha}\psi^* \\ \psi' &= e^{i\alpha}\psi \end{aligned} \qquad (11)$$

where α is a constant. As a consequence of Noether's theorem,[3] mentioned above, there is a current, j^μ, associated with the theory given by

$$j^{\mu} = -i\eta^{\mu\nu}(\psi^*\psi,_{\nu} - \psi\psi^*,_{\nu}) \qquad (12)$$

which satisfies the conservation law

$$j^{\mu},_{\mu} = 0 \qquad (13)$$

for those spinor fields, ψ, which satisfy the Dirac equation.

We can, in complete analogy with the passage from Cartesian to arbitrary coordinates, enlarge the group by letting α be a general space-time function. The spinor fields will still be assumed to transform according to equation (11). We find that the transformed Lagrangian does not retain its form under this transformation but rather adds a term of the form $\int d^4x\alpha,_{\mu} j^{\mu}$. We can compensate for this additional term by introducing a new field A_{μ} which transforms according to

$$A'_{\mu} = A_{\mu} + \alpha,_{\mu} \qquad (14)$$

and adding a term, $-\int d^4x A_{\mu} j^{\mu}$, to the Lagrangian. Then, in analogy to equation (10), we can require that A_{μ} satisfies the equations

$$F_{\mu\nu} \equiv A_{\mu},_{\nu} - A_{\nu},_{\mu} = 0 \qquad (15)$$

These equations imply that A_{μ} can always be written in the form

$$A_{\mu} = \varphi,_{\mu} \qquad (16)$$

where φ is some space-time function. Consequently,we can always perform a gauge transformation leading to a new set of potentials $A'_{\mu} = 0$ by taking for α in equations (11) and (14) the function

$-\varphi$ just as, in the relativistic case, we could always find a coordinate system in which $g_{\mu\nu} = \eta_{\mu\nu}$ for $g_{\mu\nu}$ satisfying equation (10). Again we have formally enlarged the covariance group of the theory without adding any new content to the theory. Furthermore, one can show that there is no enlargement of the conservation laws associated with the theory. This possibility of formally enlarging the covariance group of a theory from a finite parameter Lie group to one involving a number of arbitrary functions apparently always exists. Because of this possibility we see that there is not a one-to-one correspondence between the relativity principle for a given class of theories and their corresponding covariance group.

The question of the relation between the covariance group and the relativity principle of a theory has been raised since the early days of relativity theory. Kretchmann[(1)] proposed an answer which I would like to discuss briefly since it is often quoted in connection with this question and also because it is related to the role of coordinates in physics. Kretchmann said, in effect: in order to find the relativity principle associated with a given covariance group one must find out how far he can restrict the covariance group by the imposition of non-covariant restrictions on the objects appearing in the theory without, at the same time, restricting the physical possibilities admitted by the original formulation. When one has restricted the covariance group in this fashion as much as possible he will be left with some subgroup of the original covariance group. This subgroup is then defined to be the transformation group of the relativity principle.

As examples of such restrictions let me mention the gauge conditions of electrodynamics and the coordinate conditions of general relativity. In electrodynamics we can limit the gauge transformations to those of the first kind where α is a constant by imposing, for example, the Coulomb gauge condition $\nabla \cdot \underset{\sim}{A} = 0$. Kretchmann investigated to what extent it is possible to limit the group of all coordinate transformations. He proposed first a set of coordinate conditions which have lately been rediscovered by Komar[11] and used extensively by him and Bergmann in their discussion of the quantization of general relativity. These coordinate conditions are obtained by first constructing the fourteen possible scalars that can be formed using only the metric and its first and second derivatives. For a metric which satisfies the equation of general relativity in absence of matter, $R_{\mu\nu} = g^{\rho\sigma}R_{\rho\mu\sigma\nu} = 0$ all but four of these scalars are zero. The four non-zero scalars in general have different values at different space-time points except in cases where the metric has associated with it a group of symmetries or of motions. (This concept of motion has been discussed in detail in Lecture II). Except in these cases one can then use the values of the scalars at a point to serve as the coordinates of the point. These are the coordinate conditions that Kretchmann used. For the general situation there are no coordinate transformations which maintain the Kretchmann coordinate conditions. Hence he concluded that there is no relativity principle in general relativity.

Actually, Kretchmann's criterion is not a very good one for determining when a theory admits a relativity principle. For instance if this criterion is applied to special relativity, one can limit the group of Lorentz transformations to be the identity element by non Lorentz covariant restrictions. Hence one would conclude that there is no relativity principle in special relativity either. For example we can destroy the special covariance of a theory like that of Maxwell by imposing restrictions on the electromagnetic field. We could locate the origin of the reference frame by imposing conditions on the first moments of the total energy. Additional conditions could be used to fix the orientation of the axis. One can always find a Lorentz frame in which these conditions are satisfied unless the particular field we are looking at possesses some symmetry itself. In a similar manner we can fix the α in the gauge transformation (11) by requiring, for instance, that

$\psi^*(0)/\psi(0) = 1$.

What criterion, then, can we use to find the relativity principle for a given class of theories? I will try to answer this question by first comparing the situation in which we have enlarged the covariance group of a theory without changing the physical content to the situation in which we also change the physical content. In the case of the Dirac field we were able to enlarge the gauge transformations from those in which α was a constant to those in which it is an arbitrary space-time function. This enlargement brought in the new field A_μ which we then required to satisfy equations (15). This enlargement of the theory does not change the physical situation. On the other hand we could have required that A_μ satisfy the equations

$$F^{\mu\nu}{}_{,\nu} \equiv (A_{\mu,\nu} - A_{\nu,\mu})_{,\nu} = -j^\mu \qquad (17)$$

where j^μ is given by equation (12). This enlargement does change the physical content of the theory.

Similarly, we enlarged the covariance group from Lorentz to arbitrary coordinate transformations and thereby introduced the metric as an additional element to be determined. Our requirement that it satisfied equation (10) introduced no new physics since any metric which satisfies these equations is necessarily a flat metric of special relativity. We inroduce new physics by requiring the metric to satisfy the Einstein field equation.

$$R^{\mu\nu} - \frac{1}{2}g^{\mu\nu}R = - T^{\mu\nu} \qquad (18)$$

where $T^{\mu\nu}$ is the energy momentum tensor associated with the other fields and particles of the theory.

In both the electromagnetic and the gravitation cases the difference in the two extensions of the theory is apparent. In the case where we use equation (10) for the metric or equation (15) for A_μ these variables are not dynamical objects while in the case of equations (17) and (18) they are. In the first case their determination is entirely independent of the other physical objects of the theory while in the latter case this is no longer true.

In order to make more precise these differences I would like to distinguish between two different types of elements which may appear in a physical theory: absolute elements and dynamical elements. This distinction will prove important since we shall use the absolute elements of a theory to define the relativity principle associated with the theory. Let me first say how one can determine the absolute elements of a theory.

Suppose that the theory is given as a set of functional relations

$$\mathcal{L}_i(y_A) = 0 \qquad (19)$$

between the independent variables, y_A, of the theory. Furthermore suppose that equation (19) have associated with it a particular covariance group of transformations. We now look at all the invariant functions that we can form with various subsets of the y's. By an invariant function I mean one whose value does not depend upon a particular choice of gauge of coordinate system. In electrodynamics the $F^{\mu\nu}$ are invariant functions of the A^{μ}. For the group of all curvilinear coordinate transformations of general relativity, invariants are more difficult to construct. By itself, a scalar field variable is not an invariant. It becomes an invariant only if we give an invariant prescription for locating the point at which the scalar is to be evaluated. The values of these functions formed from a given subset are uniquely determined as a consequence of the equation (19) and nothing more, for example the remaining y's, boundary conditions, initial conditions, etc., then the y's that make up this subset constitute an absolute element of the theory. Of course, these y's themselves are not in general invariant under the covariance group.

The test for absolute elements is not as difficult as it first might seem. In order to know if a particular subset forms an absolute element, we need only to construct at most as many independent invariants as there are members of the subset in question. If they are all determined uniquely then any other invariants formed from the subset will also be uniquely determined since they will be functions of these original invariants. Furthermore, it will usually be quite obvious for a particular theory what subsets form invariant elements.

In the theory with $F_{\mu\nu} = 0$, the A_μ are uniquely determined up to a gauge transformation and hence any invariant formed from then is uniquely determined. Thus they form an absolute element. They do not form an absolute element when the customary Maxwell's equations (17) are assumed to hold. This is because the A's can be determined from Maxwell's equations only with the knowledge of the source currents and boundary conditions in addition to a knowledge of the gauge. Similarly when $R_{\mu\nu\rho\sigma} = 0$ the $g_{\mu\nu}$ form an absolute object since they are uniquely determined up to an arbitrary coordinate condition. But this is not the case when the g's are assumed to satisfy the Einstein equation (18).

As another example of a theory with absolute elements I will give one which was proposed in the early days of relativity as an alternate possible gravity theory. It required that the metric satisfy the equations

$$C_{\mu\nu\rho\sigma} = 0 \tag{20}$$

and

$$R = 0 \tag{21}$$

where $C_{\mu\nu\rho\sigma}$ is the conformal or Weyl tensor formed from the metric and its first two derivatives and R is the curvature scalar. It can be shown that any metric which satisfies equation (20) is conformally flat, that is, can, by means of a coordinate transformation, be made to take the form

$$g_{\mu\nu} = \gamma(x)\eta_{\mu\nu} \tag{22}$$

where $\gamma(x)$ is an arbitrary space-time function. Along with boundary and initial conditions $\gamma(x)$ is determined by equation (21). This theory possesses a spherically symmetric static Schwarzschild-like solution but it gives the wrong value for the advance of the perihelion of Mercury. If we introduce new variables $(\sqrt{-g})^{-\frac{1}{2}}g_{\mu\nu}$ and $\sqrt{-g}$ then the $(\sqrt{-g})^{-\frac{1}{2}}g_{\mu\nu}$ form an absolute element. It is interesting to note that equations (20 and 21), Einstein's equations (18) and the flat-space equations (10) are the only generally covariant, local second order equations that one can require the metric to satisfy.

Having defined the absolute elements of a theory we can now determine the relativity principle for the theory. We shall define the relativity group associated with the relativity principle as that subgroup of the covariance group of the theory which leaves the absolute elements of the theory invariant. If there are no absolute elements then the relativity group is identical with the covariance group.

If we apply this criterion to the two electromagnetic theories characterized by equations (15) and equations (17) we see that, in the former case, since A_μ is an absolute element, the relativity group is just the totality of gauge transformations of the first kind with α a constant. In the latter case A_μ is no longer an absolute element and hence the relativity group is that of all gauge transformations with α an arbitrary space-time function. Similarly, when the metric satisfies $R_{\mu\nu\rho\sigma} = 0$ it is an absolute element and the relativity group is the group of Lorentz transformations. When the metric satisfies equation (18) it is no longer an absolute element and so the relativity group is then the group of all arbitrary coordinate transformations with non-vanishing determinant. We see that with the above definition of a relativity group we obtain the expected results in each of the cases discussed.

I want to discuss the significance of relativity groups and absolute elements in physics. But first I would like to criticize the approach to "preferred" coordinate systems in general relativity* taken by Fock[(2)] and to comment on the

---

I will continue to use the term "general relativity" to describe Einstein's theory in spite of Fock's objections to the term. I believe that I have given a precise definition which makes the term meaningful.

relation of conservation laws to relativity groups.

Fock has suggested that the harmonic coordinate conditions

$$\{\sqrt{-g}\, g^{\mu\nu}\}_{,\mu} = 0 \tag{23}$$

together with certain conditions at infinity such as no incoming gravitational radiation, determine a preferred set of coordinate systems. To justify the term preferred for these systems Fock asserts with the support of plausibility arguments that, in the case of an isolated system of masses, the harmonic conditions together with suitable supplementary conditions determine the coordinate system uniquely up to a Lorentz transformation. He also points out that the harmonic coordinates satisfy a linear, generally covariant equation. Fock further argues that "Only if the existence of such a coordinate system is recognized as reflecting certain intrinsic properties of space-time can one speak of the correctness of the heliocentric Copernican system in the same sense as this is possible in Newtonian mechanics. If this is not recognized, or if the existence of the preferred coordinates is denied, one is led to the inadmissible point of view that the heliocentric Copernican system and the geocentric Ptolemaic system are equivalent."

While Fock implies that the existence of his preferred coordinates reflect some intrinsic properties of space-time he has not said what these properties are or how they are related to the harmonic coordinate systems. As far as I can see, his arguments in favor of the harmonic coordinates are of the same nature as those that might induce us to call the Cartesian coordinate systems preferred in special relativity. While it is certainly true that the use of Cartesian coordinates in special relativity simplifies many things, there is no physical reason why we cannot set up other coordinate systems. In fact we often do. For example, the hydrogen atom is best described in spherical coordinates. What is essential and physical in special relativity is the singling out of a class of reference frames, the inertial ones, from among the totality of all possible reference frames by the associated relativity principle. How we happen to coordinatize an inertial frame is of no physical significance but merely a matter of convenience. Similarly, in the case where A_μ satisfies $A_{\mu,\nu} - A_{\nu,\mu} = 0$, the gauge frame in which $A_\mu = 0$ might be preferred on the grounds of simplicity. But nothing is changed physically if we use some other gauge frame. In either case, the relativity group is the group of gauge transformations of the first kind. Only if there is some physical reason why we can only use one or another coordinate system is it meaningful to talk about a preferred system. Otherwise one is forced to use the vague criterion of "most natural" or "simplest" in picking out a preferred system.

At the beginning of this discussion I mentioned the relationship that exists between the invariance properties of a theory and the conservation laws associated with this theory. This relationship is revealed in the theorems of Emmy Neother[3]. Usually the results of the Noether theorems are given in two parts. One part applies to p-parameter Lie groups of transformations and the other to groups of transformations which depend upon q arbitrary functions of the space-time coordinates. Actually the two cases are not basically different as Bergmann[2] showed, since any group of the second kind contains an infinity of one-parameter subgroups, generated by all possible sets of the q functions.

The statement of the Noether theorem follows. We are given a theory with a relativity group (in the sense in which we have used the term) that is a p-parameter Lie group G_p and whose equations of motion for the field variables y_A are derivable from a variational principle. If ϵ^i $(i = 1,\ldots,p)$ are the parameters of the G_p, then there exists a number of quantities $t^{\mu}{}_i$ $(\mu = 1,\ldots,4)$ that satisfy p continuity equations of the form

$$t^{\mu}{}_{i,\mu} = 0 \tag{24}$$

whenever the equations of motion for the field variables are satisfied. This result only holds provided that the group G_p

is a true relativity group associated with a relativity principle and does not arise as a consequence of the introduction of absolute elements into the theory. Thus, while the group of general coordinate transformations contain an infinity of one-parameter Lie groups they do not, in general, lead to continuity equations of the form (24) if the metric is an absolute element in the theory. Only the Lorentz group leads to continuity equations in special relativistic theories. In these cases equation (24) expresses the conservation law for the stress-energy tensor.

The conservation equations (24) are, as I said, only satisfied by solutions of the equations of motion and as a consequence are sometimes called weak laws. If the theory has a relativity group whose transformations depend on a number of arbitrary space-time functions and if this group contains G_p as a subgroup then the conservation laws associated with G_p can be extended to strong laws that hold whether or not the equations of motion are satisfied. These laws are

$$\Theta^{\mu}_{i,\mu} \equiv 0 \,. \tag{25}$$

As a consequence one can infer the existence of a set of superpotentials $U^{\mu\nu}_i$ with the properties that

$$\Theta^{\mu}_{i} = U^{\mu\nu}_{i\ ,\nu} \tag{26}$$

and

$$U^{\mu\nu}_{i} = -U^{\nu\mu}_{i} \,. \tag{26}$$

In electrodynamics the superpotentials are just $F^{\rho\sigma}\xi$ where ξ is an arbitrary space-time function. Thus there exists an infinity of conservation laws

$$\Theta^{\mu}{}_{,\mu} \equiv 0 \tag{27}$$

where

$$\Theta^{\mu} = (F^{\mu\nu}\xi)_{,\nu} \ . \tag{28}$$

To date the only one of these conservation laws that can be given a simple interpretation is for the case when $\xi = 1$. Then, when the field equations are satisfied, $\Theta^{\mu} = j^{\mu}$ and we have the usual continuity equation for the current-density four-vector. It is possible to interpret some of the terms appearing in other Θ's in terms of higher electric and magnetic moments of the charge distribution but it is not clear that they lead to anything useful.

In the case of general relativity there again exists an infinity of superpotentials which in turn lead to a corresponding number of continuity equations. There are actually a number of alternate expressions for the superpotentials that differ from each other by quantities that are skew-symmetric in the upper two indices. One set of superpotentials are:

$$U^{\mu\nu} = (16\pi\sqrt{-g})^{-1} g_{\kappa\lambda}\{g(g^{\mu\delta}g^{\nu\kappa} - g^{\nu\delta}g^{\mu\kappa})\}_{,\delta}\xi^{\lambda} \tag{29}$$

where the ξ^{λ} are four arbitrary space-time functions. These superpotentials can again be used to construct conserved currents $\Theta^{\mu} = U^{\mu\nu}{}_{,\mu}$. If, in a particular coordinate system, one lets each of the ξ^{λ} take on the value unity while the others are set equal to zero he obtains four continuity equations

$$\Theta_{\mu}{}^{\nu}{}_{,\nu} \equiv 0 \tag{30}$$

where

$$\Theta_{\mu}{}^{\nu} = t_{\mu}{}^{\nu} + T_{\mu}{}^{\nu} . \tag{31}$$

Here $T_{\mu}{}^{\nu}$ is the stress-energy tensor due to matter fields, etc. while $t_{\mu}{}^{\nu}$ is the Einstein pseudotensor of stress-energy. The usual interpretation of $t_{\mu}{}^{\nu}$ is that it represents the stress-energy of the gravitational field. However it does not transform like a tensor density under arbitrary coordinate transformations and in fact is not even a geometrical object. This fact has given rise to endless discussions of the role and meaning of energy in general relativity. It seems fairly clear by now that any attempt to single out, from the infinity of continuity equations that follow from the superpotentials $U^{\mu\nu}$ of eq. (29), four special ones to describe energy and momentum conservation is doomed to failure. Only in very special cases where the metric admits a motion group is this possible.[3] The essential point is that in general relativity the relativity group of all coordinate transformations leads to an enlarged class of continuity equations as compared to the situation in special relativity. Whether or not all of these continuity equations are meaningful and can be tested in principle at least, by observation is still an open question. A definitive answer one way or the other would of course shed a great deal of additional light on the general relativity principle.

I have indicated how the relativity principle associated with a particular theory is determined by the absolute elements of that theory. I would like to conclude this disucussion with a few words about the inverse relation. It is clear that some such inverse relation must exist. Thus, if we insist that the group of all coordinate transformations is a relativity group rather than just a covariance group for the theory, we are forced to treat the metric as a dynamical, as opposed to absolute element, since no metric admits the group of all possible motions. The requirement that the group of all coordinate transformations be the relativity group of physics is thus by no means a trivial statement. If we add the requirement that the equations that determine the metric are local equations and are of second differential order in the metric then there is just one system of equations that satisfy these requirements, namely equations (18).

From what we have said, it appears that the relativity principle one assumes determines the absolute elements of the theory and at the same time greatly restricts the class of possible theories one can construct consistent with this relativity principle. We are thus supplied with a very useful tool to guide us in formulating physical theories. In particular, the relativity principle helps us single out the absolute elements in a theory. Suppose that a given theory has an obvious relativity principle associated with it such as in the case of special relativity and that the associated relativity group is a subgroup of some larger group, for example the group of all coordinate transformations. Then in general it may be possible to reformulate the original theory so that its covariance group is the larger group. However to do so we must introduce additional, absolute elements into the theory. Actually these elements were there in the first place, although their existence was masked by the fact that they had been assigned particular values. That is, the $g^{\mu\nu}$ are present in special relativity with the fixed preassigned values of the Minkowski metric. However, once we have called attention to their role as absolute elements in the theory, we can raise the general question of the validity of a theory which admits them in this role. To elaborate on this, I will discuss what might be called a "general principle of reciprocity".

It is seen that the absolute elements affect the physical behavior of a system. That is, a different assignment of values to the absolute elements would change the physical behavior of the system. Assigning different values to the metric might result in particle paths that are circles rather than straight lines. On the other hand, the physical behavior of a system does not affect the absolute elements. An absolute element in a theory indicates a lack of reciprocity; it can influence the physical behavior of the system but cannot in turn be influenced by this behavior. This lack of reciprocity seems to be fundamentally unreasonable and unsatisfactory. We may express the converse in what might be called a general principle of reciprocity: each element of a physical theory is influenced by every other element. In accordance with this principle, a satisfactory theory should have no absolute elements. It was this dislike for absolute elements that in part led Einstein to treat the metric as a dynamical element and to deduce the equations of motion (18).

What then is the role of the notion of absolute elements in a theory? First it can be used to judge a theory with regard to its satisfying the above principle of reciprocity. If it contains absolute elements, it is unsatisfactory. We then must extend the theory so that these elements become dynamical elements and the relativity group becomes the entire group of transformations, and that there are no remaining absolute elements. In doing this we can use the fact that our new equations must be covariant with respect to the enlarged relativity group of transformations to help discover the form of these equations. If further we require that these equations be local in the sense that they can be verified by purely local means then we can restrict the possible equations to a very few.

I would like to illustrate the consequence of the above discussion in terms of the justification it provides for introducing Yang-Mills[14] type fields into physics. From our point of view, these fields are always present in a theory which is invariant with respect to rotations in isospace but they are predetermined absolute elements. When we enlarge the covariance group to include the possibilities of different rotations in different directions at each space-time point, these fields appear explicitly but can be required to satisfy equations analogous to (15) which does not change their status as absolute elements. If we demand that these fields be physical elements, then we must extend the theory as Yang and Mills did. This example suggests that we should examine other transformation groups in physics to see if they can be imbeded in some larger group. Then the theory which admits the original group as a relativity group can be made to admit the enlarged group as a covariance group with the addition of new elements into the theory. We could then ask if these new elements should remain as absolute elements or if, instead, the theory should be enlarged so that the covariance group becomes a relativity group and the absolute elements become dynamical elements.

I would like to conclude this discussion of relativity groups and absolute elements with a few comments on approximate symmetries and the strong interactions of strange particles. When we speak of a symmetry of a system we refer to a particular physical situation. It is something which, in principle at least, can be observed directly. Thus we speak of the spherical symmetry of the field of a point electron. As a consequence of Noether's theorem there are a number of conserved quantities associated with this symmetry, for example, the angular momentum of a charged particle moving in a spherically symmetric field. It sometimes happens that some element always appears to possess a certain type of symmetry whenever we look at it. We tend then to say that this symmetry is a law of nature and to formulate other laws of nature so as to include it. When we do so the element with the symmetry becomes an absolute element in the theory. If we accept the hypothesis that there are in fact no absolute elements in physics we see that the observed symmetry can be explained in the framework of the theory wherein the absolute element is taken as a dynamical element by saying that it interacts only very weakly with the rest of the physical system. The symmetry it would possess in the absence of interaction is thus approximately maintained in practice. Perhaps we can look upon the conservation laws and associated symmetries of strong interactions as being due to the presence of additional elements that interact only very weakly with the strange particles in much the same way as we now think of the gravitational field, which, in the absence of matter, has the symmetries of the Lorentz group but, when allowed to interact with matter, loses this symmetry.

REFERENCES

1) E. Kretchmann, Ann. der Physik, 53, 575 (1917)

2) V. Fock, The Theory of Space Time and Gravitation, Pergamon Press, New York, 1959

3) E. Noether, Gottingen Nachr., 235 (1918)

4) H. Poincaré, R.C. Circ. Mat. Palermo, 21, 129 (1906)

5) A. Einstein, Ann. Phys. Lpz., 17, 891 (1905)

6) M. Planck, Berl. Ber., 542 (1907); Ann. Phys., 26, 1 (1908)

7) A. Einstein, Jahrb. der Radioaktivität und Elektronik, 4, 411 (1907)

8) R. C. Tolman, Relativity, Thermodynamics and Cosmology, Oxford, 1934, Chap. V

9) F. Juttner, Ann. Phys. Lpz., 34, 856 (1911), also G. E. Tauber and T. W. Weinberg, Phys. Rev. 122, 1342 (1961)

10) J. J. Sakurai, Ann. Phys., 11, 1 (1960)

11) A. Komar, Phys. Rev., 111, 1182 (1958)

12) P. G. Bergmann, Phys. Rev., 112, 287 (1958)

13) A. Trautman, Conservation Laws in General Relativity, a preprint

14) C. N. Yang and R. L. Mills, Phys. Rev., 96, 191 (1954)

THE DEGENERATE STAR,

A RELATIVISTIC CATASTROPHE

Lecture X

J. A. Wheeler

Princeton University

Seminar on Gravitation and Relativity, NASA Goddard Space Flight Center, Institute for Space Studies, New York, N. Y.; edited by H. Y. Chiu and W. F. Hoffmann.

Preface

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R.H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution the space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a certain amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and correct the final manuscript. Hence, we accept responsibility for errors and omissions.

H.Y. Chiu

W.F. Hoffmann

I. INTRODUCTION

In this lecture I shall consider the structure of stars of high density at zero temperature. For the sake of simplicity I shall assume the star has reached the end point of thermonuclear evolution and all the elements are in a state of statistical equilibrium. The question I shall ask is: Is there a limit for the mass of such stars? And, if there is, what would happen if such mass limit is exceeded?

I would like to remark that this problem stands in the very frontiers of elementary particle physics and gravitational physics. There are some interesting paradoxes which I am not prepared to give answers to. These paradoxes deal with elementary particles, on the one hand, and the geometrodynamics, on the other.

In the past this problem was connected with the determination of the proper equation of state at a zero temperature. For the sake of simplicity we shall limit ourselves to stars of zero absolute temperature. Even so, the complexity of the equation of state is still not solved. We have a general idea about the equation of state up to nuclear density, ($\sim 10^{14}$g/cm^3) but not beyond. However, as we shall show later, the proper form of the equation of state will have little to do with the general conclusion we shall arrive at.

Since we limit ourselves to the consideration of stars at zero temperature, we can consider the star to be nearly at thermodynamic equilibrium with the space surrounding it. Consequently, we can disregard radiative transfer processes inside the star. Moreover, the star is also assumed to have reached the end of its thermonuclear evolution. In almost all stars this end is far from being reached yet. The nuclear processes inside them are just on their way to convert hydrogen or helium into heavy elements. However, from nuclear physics and thermodynamics we can infer what state of affairs exist when the end of thermonuclear evolution is reached. The end point is quite independent of the intermediate nuclear processes. Of course, we shall have no knowledge of the rate at which the end point is reached, but this is not what we are interested in.

In the past the problem of the structure of stars at zero temperature has been carried out in two phases. In one phase, the density of the star considered is taken to be relatively low, being around $10^6 g/cm^3$. The star derives its pressure to counteract the crushing force of gravity from the electrons. At zero temperature the electrons are degenerate. If the mass of the star is not too high, the Fermi pressure of the electrons is enough to counteract the gravity. When the mass of the star exceeds a certain limit, (around 1.2 $M_\odot$), the Fermi pressure of electrons is not enough to counteract gravity and in principle the star collapses to a point. (For details, see Appendix.) But before the star is squashed to a point, the electrons will be crushed out of existence (being absorbed by protons to form neutrons) and the assumption that the electron gas pressure supports the star is no longer valid. Detailed theory of stars of this category, known as white dwarfs, has been furnished by Chandrasekhar a long time ago. The critical mass M_{cr} at which the theoretical radius for white dwarfs is just zero is known as the Chandrasekhar mass limit. M_{cr} is given as:

$$M_{cr} = (5.73/ \quad (\mu_e)^{2}) \quad M_\odot . \qquad (2)$$

where μ_e is the ratio of numbers of nucleons to electrons of the star. μ_e is very close to 2 for all elements except hydrogen.

When the density is very high, the electrons are crushed out of existence by the inverse beta decay process:

$$P + e^- \rightarrow n + \nu \qquad (3)$$

so that the star essentially is composed of neutrons. This is the other phase of the theory of zero temperature stars.

The pressure of the neutrons holds the star against the squashing force of gravity. There is also a mass limit for such neutron stars. The mass limit was considered by Landau first, and later by Serber. Detailed numerical calculations were done by Oppenheimer and Volkoff in 1939. They considered the neutron gas to be a perfect Fermi gas without internal structure (the internal structure of nucleons was not looked at closely until 1950), and they took into account the general relativistic effect (to be discussed below). The critical mass they obtained was 0.76 $M_\odot$, considerably smaller than that for white dwarfs.

In the following sections I shall first derive the equations of stellar structure in the absence of general relativistic effect. The more general equations as used by Oppenheimer and Volkoff (Phys. Rev. <u>55</u> 374 (1939)) shall be stated only. For detailed derivation the reader may refer to their original paper. Next I shall discuss the equation of state. Afterwards, the results of Chandrasekhar and that of Oppenheimer and Volkoff will be described. Finally, we shall review the problem that was first considered by Oppenheimer and Snyder, the problem of continued gravitational contraction when the stellar mass exceeds the critical mass for white dwarfs and neutron stars.

II. THE EQUATIONS OF STELLAR STRUCTURE

a) The equation of stellar structure at zero temperature without general relativistic correction.

Here we assume (1) spherical symmetry, (2) no rotation, (3) static equilibrium.

At every point, r, of the star there is an inward gravitational force due to the mass M(r) contained inside the sphere of symmetry of radius r. (Figure 1)

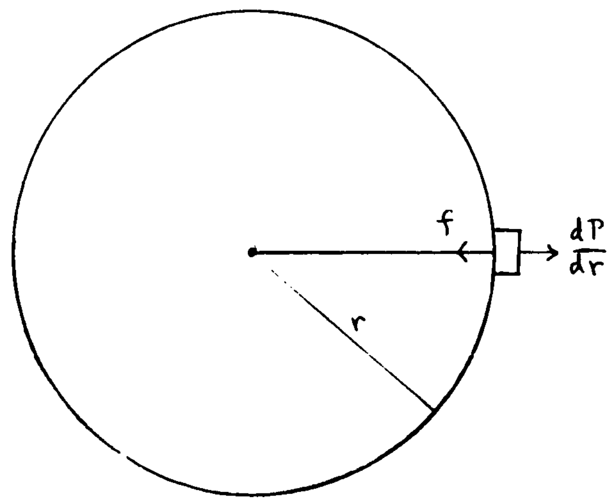


Figure 1.

The Hydrostatic Equilibrium of a Star

The gravitational force f acting on a unit volume at r is:

$$f = \rho \frac{GM(r)}{r^2} \qquad (4)$$

where G is the gravitational constant. f is balanced by the hydrostatic pressure gradient $\frac{dP}{dr}$. Therefore, the first equation of stellar structure is:

$$\frac{dP}{dr} = -\rho \frac{GM(r)}{r^2} \tag{5}$$

The second equation defines M(r):

$$M(r) = \int_0^r 4\pi\rho \; r^2 \; dr \tag{6}$$

Usually Equation (6) is written in the form of the following differential equation:

$$\frac{d\,M(r)}{dr} = 4\pi \; r^2 \rho \tag{7}$$

However, a glance at Equations (5) and (7) tell us that there are no unique solutions because there are only two differential equations to determine three dependent variables, M(r), ρ (r), and P(r). The equation of state supplies us with a relation between P and ρ :

$$P = P(\rho) \tag{8}$$

The boundary conditions for Equations (5) and (7) are as follows:

$$\text{at } r = 0, \quad M = 0, \text{ and } \quad \rho = \rho_c$$

$$\text{at } r = R, \quad P = 0, \quad \rho = \rho_o \quad \text{and } M = M_{star} \tag{9}$$

where R is the radius of the star, and M_{star} is the mass of the star. It should be noted that there is some redundancy in these boundary conditions. We shall discuss them later.

b) The equations of stellar structure at zero temperature with the general relativistic correction:

Under the same assumptions (1) (2) (3) of the last section, when the general relativistic effect is taken into account Equation (5) is modified to be:

$$-\frac{dP}{dr} = \frac{\left(\rho(r) + \frac{P}{c^2}\right) G\left[M(r) + \frac{4\pi r^3}{c^2} P(r)\right]}{r\left[r - \frac{2GM(r)}{c^2}\right]} \quad (10)$$

while Equation (7) remains unchanged. The meanings of the extra terms that appear in Equation (10) are as follows: In Equation (5) $\rho(r)$ stands for the energy density c^2. We have thus interpreted ρ in Lecture 4, in deriving the Einstein field equations. The energy density includes not only the rest energy of the mass, but also the stress energy density which is P. Therefore, it is natural to see P appearing together with ρ . Similarly the stress energy should be also included in the mass term M(r). The presence of large density causes the space to be curved. This is reflected in the denominator in Equation (10). The metric and coordinates used here are of the Schwarzschild type.

c) The Equation of State.

The equation of state

$$P = P(\rho) \qquad (8)$$

provides an additional functional relation among the dependent variables so that Equations (5) and (7) may be solved.

Equation (8) must be obtained from the statistical consideration of the kinetic properties of the gas. The pressure of the gas arises from either the electrons or the neutrons. Both are Fermions. Hence the gas pressure is essentially a degenerate Fermi gas pressure. If there is no nuclear transition, and the gas is perfect, free from any internal degrees of freedom other than spin, and the kinetic energy of the gas is small compared with its rest energy (non relativistic degenerate gas), then

$$P = K_1 \rho^{5/3} \qquad (11)$$

If the contrary is true, then

$$P = K_2 \rho^{4/3} \qquad (12)$$

K_1 and K_2 are constants depending on the mass of the gas particles. However, even in the presence of nuclear transition effect, the dependence of P on ρ is usually monotonic.

When are such simple relations as shown in Equation (11) and (12) valid? H. Harrison, M. Wakano and myself looked into this problem from the point of view of nuclear equilibrium. Salpeter and Cameron later looked into this problem with the same point of view. The idea is that under nuclear equilibrium this quantity b:

$$b = \frac{B(Z,A) - (E_F - c^2(m_n - m_H))}{A} \qquad (13)$$

should have a maximum value with respect to arbitrary changes of Z and A. $B(Z\text{-}A)$, the binding energy of the nucleus (Z, A), is given by the Weizsäcker semi-empirical formula. E_F is the Fermi energy of electrons. m_n is the mass of free neutron and m_H that for hydrogen atom. The equilibrium composition for different densities is tabulated below: (Salpeter, Astrophysical Journal 134 669 (1961)).

E_F(mev)	0.6	2.5	3.9	6.1	7.0	8.5	9.5	14.8
$\log_{10}\rho(g/cm^3)$	7.15	8.63	9.15	9.69	9.87	10.13	10.28	10.84
nucleus	(26,56)	(28,62)	(28,64)	(28,66)	(28,68)	(30,76)	(30,78)	(30,80)

E_F	20.6	24.0	11.53
$\log_{10}\rho$	11.28	11.53	
neutrons	(32,90)	(38,120)	neutrons

At low density the equilibrium composition is mainly

Fe^{56}. As the density increases the Fermi energy of electrons will increase and eventually it will be so high to promote the inverse beta process:

$$e^- + (Z,A) \rightarrow (Z-1,A) + \nu \qquad (14)$$

The threshold energy of electrons for reaction (14) is the binding energy difference of (Z-1,A) and (Z,A). As reaction (14) goes on and on the ratio of A to Z will increase. The nucleus increases in size until the number of neutrons are so high that the binding energy is essentially zero and the nucleus disintegrates into free neutrons. Just before the formation of free neutrons the final element formed is Sr^{122}.

Actually, the final composition is not pure neutrons. It will be a mixture of about 1/8 as many protons and electrons as neutrons. This is the limiting composition without considering the presence of hyperons.

The limiting ratio of eight neutrons to one proton may be understood in the following way: The equilibrium reaction among the neutrons, protons, and electrons is:

$$n \rightarrow P + e^+ + \nu \qquad (15)$$

If the Fermi energy is high, we may neglect the energy difference of the neutron to the proton-electron system (1 mev).

The neutrinos may be neglected when we talk about equilibrium configuration. The energy momentum relation in the relativistic limit is simply

$$E = cp \tag{16}$$

From Equation (15) we can infer that the energy of the neutron is roughly the sum of that for the electron and the proton and this energy is equally divided between the electron and proton. The wave length λ($\lambda \propto \frac{1}{\varepsilon}$) of the electron and the proton will be twice that of the neutron. Therefore, the spacial volume the neutron will occupy ($\sim \lambda^3$) is roughly 1/8 of that of the proton or electron. Hence it will take 8 neutrons to fill up the same volume occupied by a proton or electron when equilibrium is reached.

The result for the equation of state is presented in Figure 2 in which $\frac{P}{\rho}$ is shown as a function of ρ .

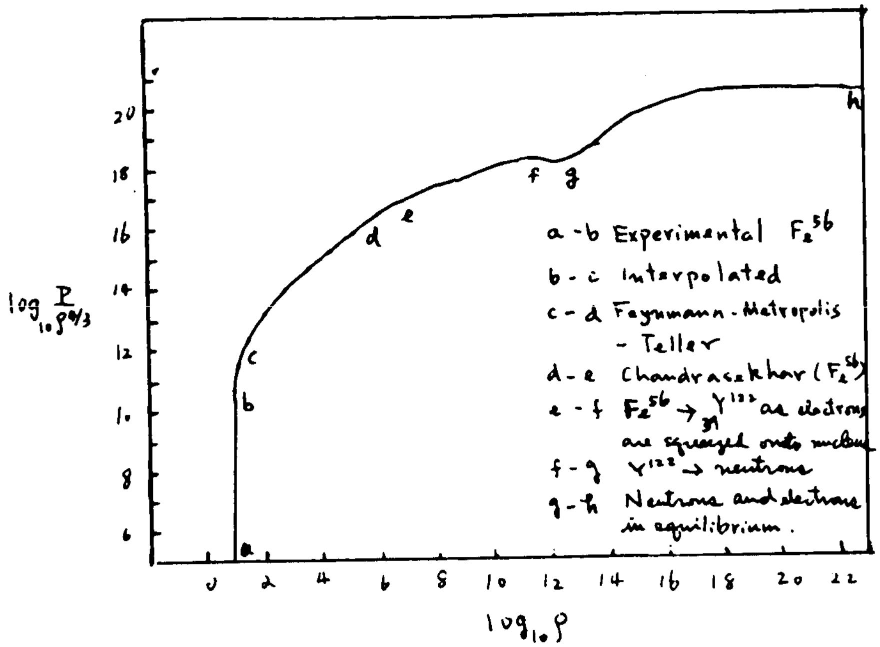


Figure 2.

The Pressure Density Relation of a Zero Temperature Gas. $\frac{P}{\rho}$ and $\frac{P}{\rho^{4/3}}$ are plotted as functions of ρ respectively, as indicated.

When the density if low, the material ia mainly made of solid iron of density 7.8 g/cm^3. It has very little compressibility. Therefore $\frac{P}{\rho^{4/3}}$ rises almost vertically until the atoms in iron are pressure-ionized. Then $\frac{P}{\rho^{4/3}}$ levels off. This is the region of atomic physics. When the pressure ionization is completed, $\frac{P}{\rho^{4/3}}$ rises as $\rho^{1/3}$ (non relativistic Fermi gas). When $\rho \sim 10^6$ g/cm^3, the electrons become relativistic and $\frac{P}{\rho^{4/3}}$

stays constant . At $\rho \sim 10^7 \rightarrow 10^{11}$ the nuclear transition occurs and $\frac{P}{\rho^{4/3}}$ drops off a little bit. After this, the composition is mainly neutrons and $\frac{P}{\rho^{4/3}}$ rises again. At $\rho \sim 10^{15}$ g/cm^3 nuclear density is reached. We do not know what happens beyond this density. We may assume that the state behaves as a hard core gas. Then $\frac{P}{\rho^{4/3}}$ should rise vertically after the density corresponding to the close packing of hard cores is reached. In this graph we have assumed the opposite, namely that the gas behaves as a perfect Fermi gas with 1/9 electrons and protons and 8/9 neutrons. There is a great deal of uncertainty to the equation of state when the nuclear density is reached.

We shall put an upper limit to this uncertainty by considering the gas as incompressible when the density is beyond 10^{15} of 10^{16}g/cm^3 -- a density corresponding to the close packing of the repulsive core of the nucleons. Any real gas is less incompressible than an incompressible gas. Therefore, this is the upper limit for any real gas. However, incompressibility is actually an absurd limit: the speed of sound in any material is governed by the ratio $\frac{P}{\rho^{4/3}}$. For an incompressible fluid the speed of sound is infinite and this violates the requirement that signals cannot be propagated with a speed greater than the speed of light. Relativity sets an upper limit 1/3 on the ratio of pressure to energy density.

If we consider the equation of state for an incompressible fluid (in spite of the above reservation), it will furnish us with a basis for analysis. We shall show that even for an incompressible fluid some drastic thing may happen to a star. Anything less than an incompressible fluid will only strengthen, and not weaken, our conclusions.

III. INTEGRATION OF STELLAR STRUCTURAL EQUATIONS

The result of Chandrasekhar and that of Oppenheimer and Volkoff.

We shall sketch a method to integrate Equations (5) or (10), together with Equation (7). The boundary conditions are as described in Equation (9).

We start by asserting a value ρ_c for the central density. Once this is done, the pressure at the center P_c is determined by the equation of state. $M(r)$ is determined in the vicinity of r, with the condition $M(o) = 0$. This condition is mandatory in order that singularities may not occur for P and ρ at the center.

$$M(r) = 4/3 \ \pi r^3 \rho_c \qquad \text{for } r \sim o \tag{17}$$

At some distance Δr from the center, the decrement of P, Δ P, may be computed from Equation (5) or (10, the pressure equilibrium condition. The decrement of density, $\Delta\rho$ is obtained from the equation of state. At some further distance, say 2 Δ r, from the center, M(r) may then be determined using $\rho + \Delta\rho$ for the density. This, in turn, determines Δp at $r = 2 \Delta r$. $\Delta\rho$ is then obtained, from which a new value of M(r) may be computed. And so on. We continue this process until P becomes zero. This occurs at some distance r = R. R is then the radius of the star. The mass of the star is given by Equation (6). Both R and M_{star} are not pre-determined. Only the central density, ρ_c , is pre-determined. Figure 3 illustrates this method of integration of a star.

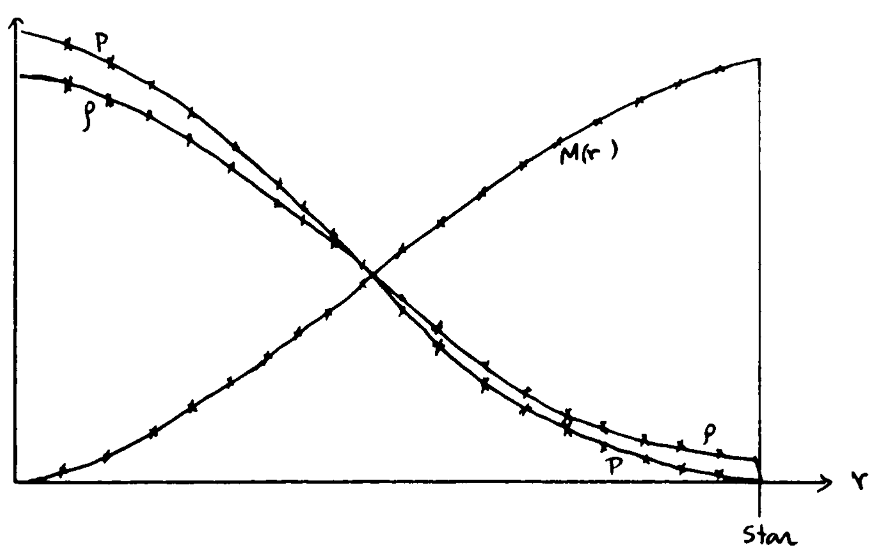


Figure 3.

Schematic Illustration of

Numerical Integrations of a Star

At $r = R$, ρ is not necessarily zero. For example, at low enough density (say $\rho \lesssim 10^8$ g/cm^3), the end product of thermonuclear evolution is iron. If the outer shell of the star has evolved to this end, the edge of the star is solid iron which has a density of 7.8 g/cm^3. The star will look like a polished iron sphere. This is the reason why in Equation (9) the density ρ_0 at which $P = 0$ is not set to zero.

In this approach we have a free choice of only one boundary condition, the value of central density.

Instead of starting with a known mass and radius of the star, and determining the central density from the stellar structure equastions, we assume a central density and calculate the consequent mass and radius. This is a very odd approach in physics. We solve the problem first before we know what the problem is. But this approach is convenient in our subsequent discussions.

Now I shall describe the result of Chandrasekhar and that of Oppenheimer and Volkoff. Let us start with a reasonably small central density, say 10^3 g/cm^3. Half of the star is pressure ionized, and the star is supported partially by the pressure of bound electrons and partly by that of the free electrons. Such an object has a mass around .01 $M_\odot$ and is something in between a star and a planet. As the central density is increased, the pressure ionization becomes more and more complete. The mass also increases. At ρ_c = 10^5 g/cm^3, the mass is around a few tenths of $M_\odot$. At ρ_c = 10^7 g/cm^3 the mass of the star is around $1M_\odot$. At ρ_c = 10^8 g/cm^3 the mass of the star is around 1.2 $M_\odot$.

This density is already very high and the nuclear transition discussed in Section 2(c) begins to take place. The number of electrons is rapidly decreasing. But let us ignore all the nuclear transitions and let us assume the number of the electrons does not change, and that the pressure is entirely due to electrons. With this assumption when ρ_c is increased without limit, M_{star} approaches an asymptotic value -- the Chandrasekhar mass limit, which, for a composition of iron is around 1.3 $M_\odot$. The dotted curve of Figure 4 shows Chandrasekhar's result which he obtained under the assumptions that the number of electrons does not change and that the pressure is entirely due to the electrons.

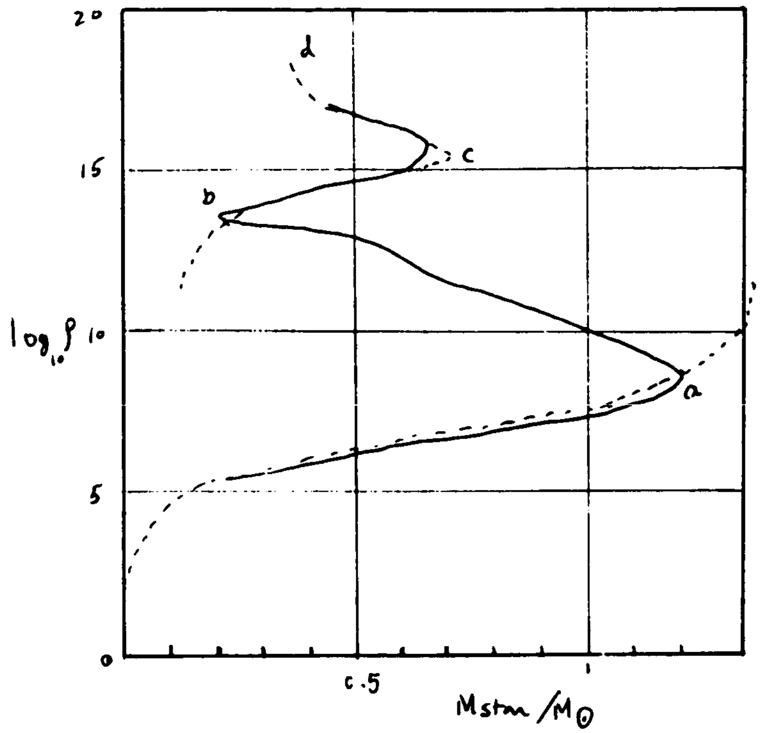


Figure 4.

The Central Density and Mass Relation for a Zero-Temperature Star

Nuclear transitions begin at a density of around 10^7 g/cm^3 and become important when $\rho > 10^8$ g/cm^3. The general relativistic effect becomes important when $\rho > 10^{11}$ g/cm^3. The solid curve of Figure 4 shows our result. For $\rho < 10^8$ g/cm^3 the solid curve and the dotted curve almost coincide. As ρ_c approaches 10^8 g/cm^3, the electrons are gradually crushed out of existence and our curve begins to depart from that of Chandrasekhar's. Instead of approaching 1.3 $M_\odot$ asymptotically, the mass reaches a maximum and begins to decrease with increasing density. I shall call this point of maximum mass (indicated as "a" on

Figure 4) the first crushing point. The part of the curve that was computed by Chandrasekhar represents the stable situation, that is, it is stable against small oscillations. If we apply a pressure pulse to this star, it oscillates with finite amplitude and ultimately it settles down to its original equilibrium, after the oscillation is damped.

However, a star which lies along the portion of the curve just above the first crushing point is not stable. If it is disturbed by the slightest amount, it departs more and more from the calculated quilibrium. Such a star is in an unstable equilibrium. The reason for this is that on curve segment a b the equilibrium mass decreases with increasing density. Physically, this is the density regime where the electrons are being squashed out of existence and $P/\rho^{4/3}$ decreases with increasing density (Figure 2). If the star is disturbed by a slight amount, some electrons undergo inverse beta decay. The pressure decreases and the star collapses a little bit, hoping to restore its initial central pressure. But the pressure does not increase rapidly enough with increasing density. So the star collapses more. This continues until the neutron pressure becomes dominant so that $P/\rho^{4/3}$ rises again.

Oppenheimer and Volkoff computed the structure of such stars under the assumptions that the gas is composed entirely of neutrons. Their result is the broken curve (b c d). They found that as the central density ρ_c increases without limit, M_{star} approaches an asymptotic value of 0.76 $M_{\odot}$. This value of the mass is smaller than that given by Chandrasekhar for white dwarfs of iron composition (1.2 $M_{\odot}$).

Now I would like to raise two questions. First, what will happen to a star if its mass exceeds the Oppenheimer-Volkoff limit? Second, how will the Oppenheimer and Volkoff result be modified by a more realistic form for the equation of state Equation (8)?

Unfortunately, we do not know the equation of state for a gas whose density exceeds 10^{15} g/cm^3, the nuclear density. However, it is useful to take the extreme form of the equation of state, that is, the equation of state for an imcompressible fluid. As we have remarked previously, the idea of an incompressible fluid is not even consistent with relativistic concept. But an incompressible fluid is an upper limit for real gases. If a star composed of an incompressible fluid collapses, then it should certainly do so for any compressible fluid.

In the next section I shall discuss a star composed of an incompressible fluid near the Oppenheimer-Volkoff mass limit.

IV. THE GRAVITATIONAL COLLAPSE OF AN INCOMPRESSIBLE STAR

As in Section III, we first consider a star of low mass so that the general relativistic effect is not important. The gravitational potential at r inside such star is given by:

$$\varphi_{in} = -G \frac{M(r)}{r} \tag{18}$$

where M(r) is the mass enclosed inside a sphere of symmetry of radius r (Figure 1). Since the star is assumed to be incompressible,

$$M(r) = \frac{4\pi}{3} \rho r^3 \tag{19}$$

where ρ is the density of the incompressible medium. Hence, inside the star:

$$\varphi_{in} = \varphi_o - \frac{4\pi}{3} G\rho r^2 \tag{20}$$

and outside the star:

$$\varphi_{out} = -G \frac{M_{star}}{r} \tag{21}$$

φ_o is determined by the condition that φ_{in} and φ_{out} must be continuous at the surface of the star and that $\varphi_{out} = 0$ at $r = \infty$.

The hydrostatic equation (Equation (5)) takes the following form:

$$- \frac{dP}{dr} = -\rho\frac{\partial\varphi}{\partial r} = \frac{8\pi}{3} G\rho^2 r \qquad (22)$$

The pressure is obtained by integrating Equation (22):

$$P = P_o - \frac{4\pi}{3} G\rho^2 r^2 \qquad (23)$$

P_o is determined by the condition that $P = 0$ at the surface of the star.

Figures 5 and 6 show the behavior of P and φ as functions of r. φ_{in} has the form of a simple harmonic potential. The set of curves (1), (2), and (3) correspond to different assumed masses for the star. In general, with increasing stellar mass, the radius, the pressure at the center, and the absolute value of φ at the center are increased.

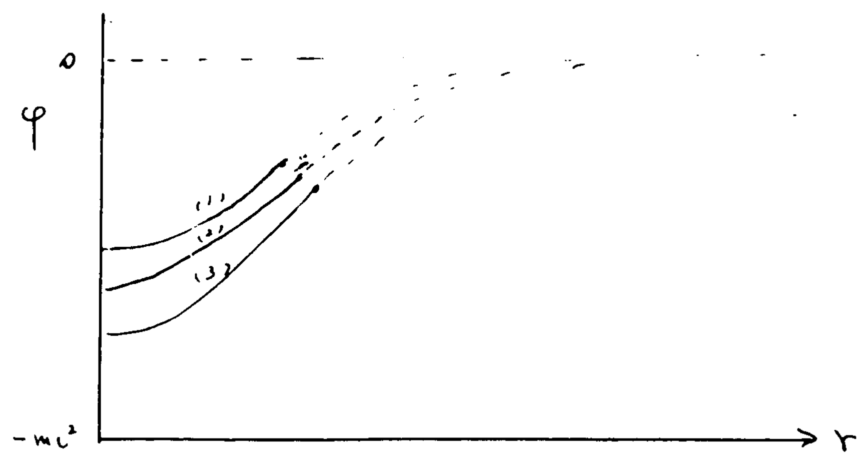


Figure 5

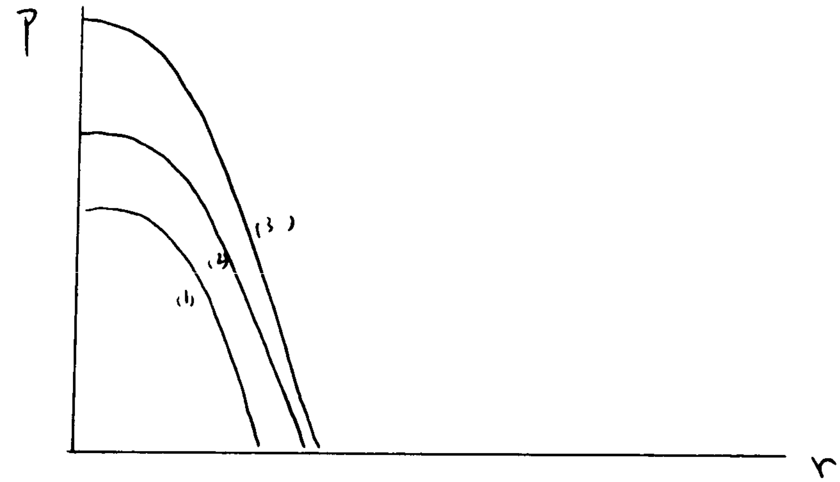


Figure 6

Now we will consider an electron-positron annihilation at the center of this star. If there were no gravitational field, the energy of the photons would be $2m_ec^2 \approx 1.02$ mev. Because of the gravitational field, the energy of the photons as measured by a distant observer will be:

$$E = 2m_ec^2 - 2m\varphi_* \qquad (24)$$

The energy required to materialize an electron-positron pair in a space free from a gravitational field is $2m_ec^2$. The gravitational field diminishes the externally supplied energy for materialization to $2m_ec^2 - 2m\varphi_*$. This energy as a function of the distance from the center is plotted in Figure 7. The numbers on the curves refer to increasing stellar masses.

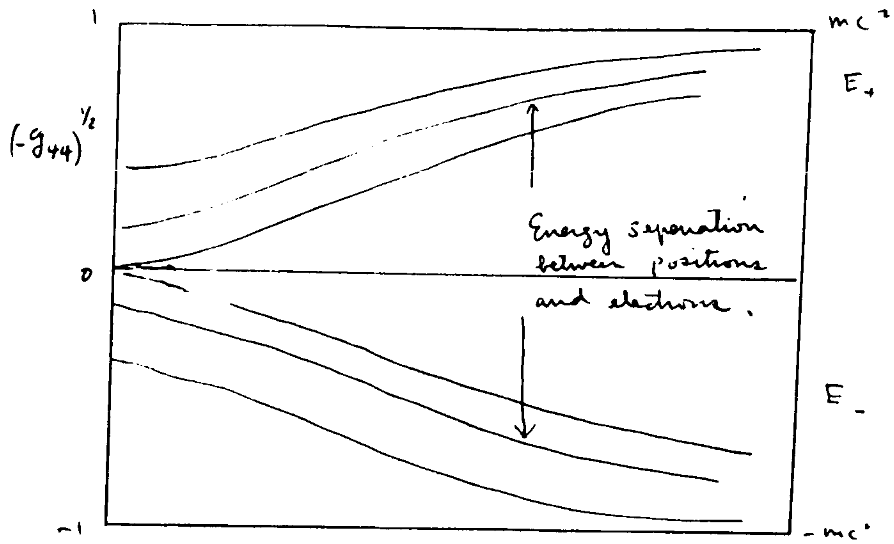


Figure 7

So far we have used only non-relativistic concepts. In the language of general relativity, the gravitational potential φ is replaced by the 4-4 component of the metric tensor $g_{\alpha\beta}$. In general, the rest energy of a particle is given by:

$$E_{\pm} = \pm\ m(-g_{44})^{\frac{1}{2}} \tag{25}$$

where m is the rest mass in the absence of gravitational field. The positive sign refers to particles (positive energy state) and the negative sign refers to antiparticles (negative energy state).

Outside the star g_{44} is given by (Lecture 4):

$$g_{44} = -(1 - \frac{2GM}{rc^2}) \tag{26}$$

For a given density ρ, there exists a mass M such that $g_{44} = 0$ at the surface of the star. This defines the critical mass:

$$M_{cr} = \frac{4\pi}{3}(\frac{3c^2}{8\pi G\rho})^{3/2} \tag{27}$$

In the above expression the radius of the star is equal to the circumference divided by 2π. If the mass of the star does reach this critical value, at the surface of this star the annihilation of a pair of electrons will not yield any energy as seen by an observer outside the star. Similarly no external energy is required to create a pair of electrons on the surface of this star.

Is this possible? How does the star ever get to this stage? What happens to matter inside?

The state $M_{star} = M_{cr}$ is certainly an idealization. By physical arguments one can prove the star cannot possess a mass $> M_{cr}$. Starting with a fluid of a given density, and a mass a little less than the critical mass, what will one observe as matter is gradually added to the star in such a way that at each stage the heat energy derived from the gravitational work is removed? For each gram added, a fraction will go towards increasing the mass of the star and a fraction will be radiated away. The radiated fraction will approach unity as the mass of the star approaches the critical mass, and very little will go into increasing the mass of the star. (The mass of stars is defined by the gravitational field they produce.)

I have made a computation of the rate of increase of M_{star} with respect to the mass (M) added. The result is:

$$\frac{dM_{star}}{dM} = \left[1 - \left(\frac{M}{M_{cr}} \right)^{2/3} \right]^{\frac{1}{2}} \qquad (28)$$

and $\frac{dM_{star}}{dM} = 0$ as $M \rightarrow M_{cr}$

This is just what we expected. However, even the limit $M = M_{cr}$ cannot be approached.

The reason for this is that we have based our arguments on the assumption that the zero separation of the positive and negative energy states occur at the surface of the star. Yet the critical condition of zero separation is already attained at the center of the sphere when the stellar mass M is 16% short of M_{cr}. For $M = 0.8382\ M_{cr}$ $(-g_{44})^{\frac{1}{2}}$ is zero at the center of the star. A value of M greater than $0.838M_{cr}$ leads to a singular and physically unacceptable behavior of the metric, in the sense that $(-g_{44})^{\frac{1}{2}}$ becomes negative for some values of r. For $M = 0.8382M_{cr}$ the pressure is given by:

$$\frac{3P}{\rho c^2} \approx 4 \left(\frac{3c^2}{8\pi G\rho} \right) \frac{1}{r^2} \tag{29}$$

$$r \ll \left(\frac{3c^2}{8\pi G\rho} \right)^{\frac{1}{2}} \equiv \text{critical "radius" of the star.}$$

As M approaches $0.8382M_{cr}$, $P \to \infty$ at the center of the star.

What happens if we allow the star to be compressible? The density will increase and by Equation (27) M_{cr} will be decreased. Hence, the critical state of vanishing g_{44} at the center will be reached for a lower masson. Starting from simple physical assumptions of the composition and the structure of the star, we have arrived at a seemingly unacceptable physical situation. We have come to the untamed frontier between elementary particle physics and general relativity.

Of all the implications of general relativity for the structure and evolution of the universe, the question of the fate of great masses of matter is one of the most challenging. The issue cannot be escaped by appealing to stellar explosion or rotational instability, for this issue as it presents itself is one of principle, not one of observational physics.

Perhaps one original assumption about hydrostatic equilibrium is not realized: This is the proposal of Oppenheimer and Snyder. They consider a collection of particles separated from their common center by distances of the order of the solar radius. They note that the fall towards the center of the star will take only a few hours as measured by an observer on one of the particles, but will take forever as measured by a remote stationary observer. They suggest the same indefinitely prolonged fate for a star whose mass exceeds the critical mass. This approach does not give an acceptable answer to the fate of a system of A-nucleons under gravitational forces for the following reasons:

(1) No mechanism for release of the gravitational energy
QUASARS
into the surroundings is taken into account. The particles are considered to convert gravitational potential energy into kinetic energy, but not into heat and radiation. Therefore

this approach excludes from the start any decrease in the mass energy of the system and rules out a priori any approach to an equilibrium, if there is one. The mass of the system as viewed by a distant observer remains forever the same. This is contrary to the physical situation in which the particles will collide, give off heat, lose speed and thus slow down their contraction.

(2) The particles are envisaged as falling into a Schwarzschild singularity. However, a Schwarzschild singularity does not give an adequate representation of the forces sustained by a particle at high compression. The forces between nucleons enter in a most vital way. Of course it is not clear what consequences these forces lead to. We have noticed that hard core forces at the one extreme of an incompressible liquid are as incapable of sustaining the system as are the pressures of a perfect Fermi gas at the opposite extreme. But either extreme is inadequate because any answer is incomplete that does not deal with the ultimate constitution of a nucleon.

(3) The particles are envisaged as "cutting themselves off from the rest of the universe" by falling inside the Schwarzschild singularity. This expression would seem to sug-

gest that the particles lose their effect on the rest of the universe. However, in Oppenheimer and Volkoff's discussions they implied that at the same time they maintain an unchanged gravitational pull on a distant test mass--the direct opposite to "cutting themselves off from the rest of the universe."

For the above reasons the Oppenheimer-Volkoff approach does not relieve us of the difficulties concerning the fate of massive stars. (It appears that the final mass of the star must be very substantially smaller than its original mass.) It must be finite, and limited to a fixed upper bound, no matter how many nucleons are in the star.

If we are to reject as physically unreasonable the concept of an indefinitely large number of nucleons in equilibrium in a finite volume of space, it seems necessary to conclude that the nucleons above a certain critical number convert themselves to a form of energy that can escape from the system as radiation. If the energy were to escape in the form of particles, we could in principle extract the energy from the emerging particles and then let them fall back on the system. The build up of these particles on the system would then ultimately lead back to the paradoxical situation from which an escape is sought. Radiation presents no such difficulty. However low its energy, it can always escape from the system by travelling radially outwards. No escape is apparent except to assume that the nucleons at the center of a highly compressed mass must necessarily dissolve away into radiation--electromagnetic, gravitational, or neutrino, or some combination of the three--at such a rate or in such numbers as to keep the total number of nucleons from exceeding a certain critical number.

In view of the absence of any acceptable alternative equilibrium, it appears desirable to take seriously this possibility of nucleonic disruption and explore its consequences. Dissolution of nucleons into neutrinos at very high pressures would be a process fully compatible with the laws of the conservation of momentum and energy. It would violate the law of conservation of nucleon number, but leave unaffected most other conservation laws. This possibility does not contradict the present lower limit to the life time of the nucleon against spontaneous decay. ($T_{\frac{1}{2}} = 4 \times 10^{23}$ years) This decay rate is determined for essentially free particles, whereas, we are dealing with nucleons in a highly compressed state.

A motion picture of a large mass of nucleons dissolving away under high pressure into free neutrinos presents a fantastic scene when run backwards. Sufficiently many neutrinos of the right helicity coming together from all directions into one region of space over a short time interval materialize into nuclear matter.

APPENDIX

The Mass-Radius Relation and the Mass Limit for White Dwarfs

We shall demonstrate here that, when the mass of a white dwarf exceeds a certain limit, no equilibrium configuration exists.

The mean density $\bar{\rho}$ is:

$$\bar{\rho} \propto \frac{M}{R^3} \qquad \text{(A-1)}$$

where M is mass of the white dwarf and R is its radius. The average gravitational force f inside the star is then:

$$f = \overline{\rho \frac{GM_r}{r^2}} \propto \frac{M^2}{R^5} \qquad \text{(A-2)}$$

The pressure P of the degenerate electron gas which supports the star has the following dependence on ρ:

$$P \propto \rho^{5/3} \propto \frac{M^{5/3}}{R^5} \quad \text{non-relativistic gas} \qquad \text{(A-3)}$$

$$P \propto \rho^{4/3} \propto \frac{M^{4/3}}{R^4} \quad \text{relativistic gas} \qquad \text{(A-4)}$$

The different dependence of P on ρ is due to the fact that in the non-relativistic approximation $E \sim \frac{P^2}{2m_e}$ and in the other $E \sim Pc$. (For details, see M. Schwarzschild, "Structure and Evolution of the Stars", p.57) From Equations (A-3) and (A-4) the average pressure gradient $\frac{dP}{dr}$ is:

$$\frac{dP}{dr} \propto \frac{M^{5/3}}{R^6} \qquad \text{non-relativistic} \qquad \text{(A-5)}$$

$$\frac{dP}{dr} \propto \frac{M^{4/3}}{R^5} \qquad \text{relativistic} \qquad \text{(A-6)}$$

For a star in equilibrium, Equation (5):

$$\frac{dP}{dr} = -\rho \frac{GM_r}{r^2} \qquad (5)$$

must be satisfied. Now we compare Equations (A-5) and (A-6) with Equation (A-2) we find that for the non-relativistic case $\frac{dP}{dr}$ and f depend on different powers of R. Thus the star has the ability of bringing the two forces into balance by adjusting its radius. For example, if $\frac{dP}{dr}$ is bigger than f, the star will expand, increasing its radius until $\frac{dP}{dr}$ and f are equal.

This is not so in the relativistic case. $\frac{dP}{dr}$ and f depend on the same power of R. Hence the star does not have the ability to achieve equilibrium by adjusting its radius. On the other hand, $\frac{dP}{dr}$ and f have different power dependence on M. Hence there exist a specific mass value, the limiting mass, for which the two forces are in exact balance. For mass greater than the limiting mass the gravitational force will always exceed the pressure force, whatever the radius. Thus the star has to collapse.

On the other hand, when the mass is smaller than the initial mass, the gravitational force will be smaller than the pressure force and the star will expand. In this expansion the density will decrease until, at least in the outer portions, the degeneracy changes from relativistic to non-relativistic. Now with increasing radius the pressure force decreases faster than the gravitational force so that eventually the two forces will come into balance.

We may conclude that in stars heavier than the limiting mass the force of the degenerate pressure is never sufficient to balance gravity, that a star lighter than the limiting mass is able to balance gravity with degenerate pressure, and that to achieve this balance in the latter case the star has to adjust its radius to the value prescribed by the mass-radius relation for degenerate models.

GRAVITATION AND LIGHT

N68-14308

Lecture XI

Joseph Weber

University of Maryland

Seminar on Gravitation and Relativity, NASA Goddard Space Flight Center, Institute for Space Studies, New York, N. Y.; edited by H. Y. Chiu and W. F. Hoffmann.

Preface

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R. H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution the space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a minimum amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and correct the final manuscript. Hence, we accept responsibility for errors and omissions.

H. Y. Chiu

W. F. Hoffmann

1. <u>Gravitational Deflection of Light by a Non Relativistic Method</u>

The idea that light interacts with a gravitational field originated more than a century ago. In 1905 Von Soldner[1] considered the deflection of light by the sun's gravitational field from the point of view of the corpuscular theory, and Newton's laws of motion.

Consider the photon as a particle of mass m ~~it turns~~ Since ~~out that~~ m is can~~celled out~~ in the equation of motion ~~so that~~ We ~~do~~ not need concern ourselves with the ~~to worry about what~~ value of m we ~~must~~ assign to a photon. ~~It~~ Let the photon pass~~es~~ through the vicinity of a larger mass M as shown in Figure 1. Let the impact parameter be R. We use rectangular coordinates, such that at infinity the path of light is parallel to the X-axis and the deflection of light occurs in the X-Y plane. The equation of motion is:

$$\frac{md^2y}{dt^2} = -\frac{GMm}{r^2}\frac{y}{r} \tag{1}$$

where

$$r^2 = x^2 + y^2 \tag{2}$$

If the deflection is small, then $y \approx R$. We write $x = ct$. With these substitutions equation (1) may be immediately integrated to give:

$$\frac{dy}{dx} = \frac{GMx}{(c^2R)}\frac{1}{(x^2+R^2)^{1/2}} \tag{3}$$

(1) Berliner Astronomisches Jahrb. - 1804 S. 161.

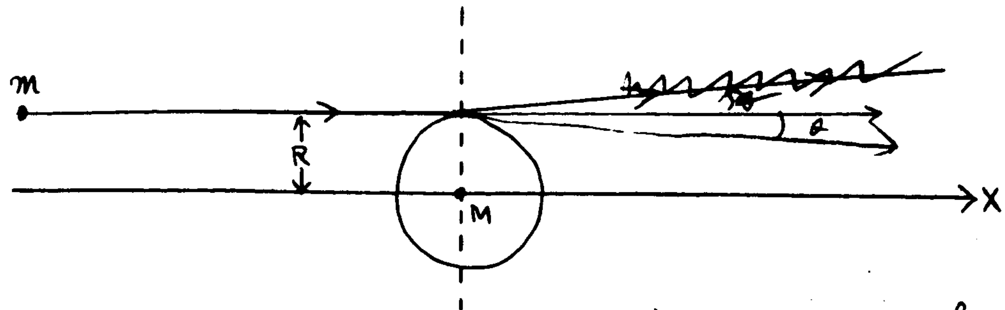


Figure 1. Deflection light by the sun in corpuscular theory.

The angle of deflection ϑ ~~of the path of light~~ is:

$$\theta \cong \left(\frac{dy}{dx}\right)_{x=-\infty} - \left(\frac{dy}{dx}\right)_{x=+\infty} = -\frac{2GM}{Rc^2} \qquad (4)$$

This is one half the value predicted by Einstein. ~~and~~ confirmed by astronomical observations during total solar eclipses. The disagreement in the two theories occurred because Von Soldner used the wrong equation of motion. His equation is valid for slow particles, but photons are not slow. We shall derive Einstein's result in the following paragraph:

2. Relativistic Results

The geodesic equation ~~is~~

$$\frac{d^2y}{ds^2} + \Gamma^{y}{}_{\mu\nu}\frac{dx^{\mu}}{ds}\frac{dx^{\nu}}{ds} = 0 \qquad (5)$$

~~For light all intervals are null.~~ s is a parameter along the path. We may choose $s = x^0 = ct$. $\Gamma^{\alpha}{}_{\beta\gamma}$ is the Christoffel symbol of the second kind. If we use isotropic coordinates, ds^2 for the lowest order of departure from a Minkowskian line element is given as:

$$-ds^2 \cong \left(1+\frac{2GM}{rc^2}\right)(dx^2+dy^2+dz^2) - \left(1-\frac{2GM}{rc^2}\right)c^2dt^2 \qquad (6)$$

With this metric $\Gamma^{y}{}_{0x} = 0$, and to a good approximation Equation (5) becomes:

$$\frac{d^2 y}{dx^2} + \Gamma^{y}{}_{00} + \Gamma^{y}{}_{xx}\frac{v^2}{c^2} = 0 \qquad (7)$$

We have neglected terms involving the y component of velocity since it is very small. From the definition of $\Gamma^{\alpha}{}_{\beta\gamma}$, and Equation (6), we have:

$$\Gamma^{y}{}_{00} = -\frac{1}{2}\frac{\partial g_{00}}{\partial y} = \frac{GM}{r^2 c^2}\frac{\partial r}{\partial y} = \frac{GMy}{c^2 (x^2 + y^2)^{3/2}} \qquad (8)$$

$$\Gamma^{y}{}_{xx} = -\frac{1}{2}\frac{\partial g_{xx}}{\partial y} = \frac{GM}{r^2 c^2}\frac{\partial r}{\partial y} = \frac{GMy}{c^2 (x^2 + y^2)^{3/2}} \qquad (9)$$

where $\frac{\partial r}{\partial y} = \frac{y}{r}$. The deflection of light by a mass center may then be computed:

$$\theta = \left(\frac{dy}{dx}\right)_{-\infty} - \left(\frac{dy}{dx}\right)_{+\infty} = -4\frac{GM}{rc^2} \qquad (10)$$

It is easy to see why we get an extra factor of 2. The deflection is given by the contribution of two identical associated with the gravitational field terms $\Gamma^{y}{}_{00}$ and $\Gamma^{y}{}_{xx}$. ~~each of which is equivalent to equation (7)~~. In the classical theory, only one term is present ($\Gamma^{y}{}_{00}$). At low velocity $\Gamma^{y}{}_{xx}\frac{v^2}{c^2} << [illegible]$. At v = c both contribute equally. We can say the photon acts as if it has a gravitational mass twice its inertial mass. ~~in this case.~~

3. Effect of the Photon Spin

Although in (2) we have taken into account the relativistic effect, we have not taken into account the spin of the photon. The equations of motion for spinning particles are somewhat different from the equations for spinless particles. I shall only state the results.[2] The angle of deflection is:

$$\theta = -\frac{4GM}{Rc^2}\left(1 - \frac{\lambda}{R}\right) \tag{11}$$

if the spin of the photon is perpendicular to its direction of motion. λ is the wavelength of the photon. If the spin of the photon is parallel to its direction of motion the extra term $\frac{\lambda}{R}$ drops out. However, from quantum field theory we know that if the particle has zero mass, then its spin must be parallel or antiparallel to its direction of motion; there are only two spin states for such a particle: full spin ahead and full spin backward. Therefore, by this coincidence, the effect of the spin disappears.

4. Gravitational Red Shift

This effect of gravitation on light has been discussed many times. However, for completeness let me include a few

---

(2) Papapetrou and Corinaldisi, Proc. Roy. Soc. A209, 259, (1951)

words about the red shift. In the case of the deflection of light by the sun, the naive argument (Newtonian Mechanics) does not lead to correct results even in the first order. But for the red shift the most naive argument does lead to correct results in the first order, using only the principle of equivalence and the doppler shift law.

The principle of equivalence states that

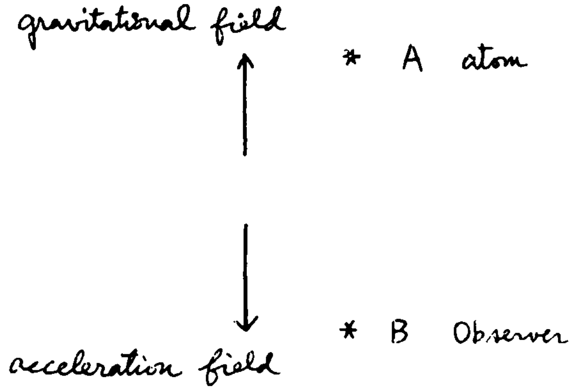


Figure 2

insofar as local observations are concerned, the effects of a uniform gravitational field are indistinguishable from those of an equivalent uniform acceleration field.

In Figure (2) the atom at A emits a photon of frequency ν. At some later time $t = \frac{\ell}{c}$ (where ℓ is the distance between the observer and the atom) the observer at B in the equivalent accelerated frame detects the photon. Over this time interval, t, the observer has undergone an acceleration of g and at the end of this time interval has attained a velocity, v.

$$v = gt = g\frac{\ell}{c} \tag{12}$$

In the gravitational field we say this gives rise to the red shift, in the equivalent accelerated frame we call it a doppler shift. The change in frequency is therefore:

$$\Delta\nu = -\nu\left(g\frac{\ell}{c}\right)\frac{1}{c} = -\nu\frac{\Delta\varphi}{c^2} \tag{13}$$

Here $\Delta\varphi$ is the change of gravitational potential, the potential at the location of the observer minus the potential at the location of the atom. (13) is the first order gravitational red shift formula.

5. Comparison of the theories of Maxwell and Einstein

In the special theory of relativity the Maxwell equations are

$$\Box A_\mu = -j_\mu \tag{14}$$

where A_μ is the 4-potential and j_μ is the current density. There is a supplementary condition to Equation (14):

$$A^\mu{}_{,\mu} = 0 \tag{15}$$

Equation (15) is the Lorentz gauge condition. On the other hand, the Einstein field equations are:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = KT_{\mu\nu} \tag{16}$$

The two sets of field equations, (14) and (16), appear quite different. In classical electromagnetic theory, equations (14) describe the electromagnetic field in Minkowski coordinates. Charges and fields operate in a flat space, in the Maxwell theory. In general relativity, the gravitational field is a property of space and is characterized by the geometry of the space. The presence of mass affects the geometry. Motions of spinless particles subjected to gravitational interactions are given by the geodesics in such a space. It was therefore believed that general relativity is fundamentally a different theory from electromagnetic theory. In recent years this view has somewhat changed.

In a more general treatment of field theories, there is a correspondence between the spin of the particle and the rank of the tensor which describes the field. In the case of the electromagnetic field, the photon has a spin 1, and a four vector is employed. It was shown by Pauli and Fierz that a second rank symmetric tensor field is required for relativistic wave equations for particles of spin 2. The reason for this is the following. A particle of spin 2 has five quantized spin orientations. Also there are two signs for the energy corresponding to particle states and anti-particle states. Hence we have a 10-component wave function for a spin 2 particle, which corresponds to the components of a second rank symmetric tensor.

Pauli and Fierz also showed that the relativistic wave equations for free, massless particles of spin 2 are

$$\Box U_{\mu\nu} = 0 \tag{17}$$

where $U_{\mu\nu}$ is the second rank symmetric tensor. The supplementary condition was given by them[3] to be:

$$U^{\mu\nu}{}_{,\nu} = 0 \tag{18}$$

Suppose now that we introduce interactions. These may be represented by some tensor $\theta_{\mu\nu}$. The field equations with interactions are then:

$$\Box U_{\mu\nu} = k\theta_{\mu\nu} \tag{19}$$

Here k is a coupling constant. The supplementary condition, Equation (18), now implies that:

$$\theta^{\mu\nu}{}_{,\nu} = 0 \tag{20}$$

Since the matter stress tensor $T_{\alpha\beta}$ satisfies Equation (20), we might associate $T_{\alpha\beta}$ with $\theta_{\mu\nu}$. The appearance of the stress tensor is to be expected in developing a theory of gravitation, because stress and energy are the source of a gravitational field. The gravitational field itself is expected to contribute some stress and energy as well. We

---

(3) In this section we raise and lower indices by the Lorentz metric, i.e., $A^{\mu} = \delta^{\mu\nu} A_{\nu}$

use $t_{\mu\nu}$ to denote such stress energy. Hence the field equations are

$$\Box U_{\mu\nu} = k(T_{\mu\nu} + t_{\mu\nu}) \tag{21}$$

If $T_{\mu\nu} = 0$, the vacuum equations are obtained as:

$$\Box U_{\mu\nu} = kt_{\mu\nu} \tag{21a}$$

Now we shall try to obtain Equation (21a) from a variational principle. First we construct a Lagrangian density L for the field. The method of constructing L is described in standard textbooks on field theory.[4] The Lagrangian density for a free field is:

$$L = -\frac{1}{2} U^{\mu\nu}{}_{,\alpha} U_{\mu\nu}{}^{,\alpha} \tag{22}$$

The action principle leads to the field equations

$$\frac{\partial}{\partial x^{\alpha}} \frac{\partial L}{\partial U^{\mu\nu}{}_{,\alpha}} - \frac{\partial L}{\partial U^{\mu\nu}} = 0 \tag{23}$$

An object constructed from L and satisfying the conservation law $t_{\mu}{}^{\nu}{}_{,\nu} = 0$ is[5]

$$t_{\mu}{}^{\nu} = \delta_{\mu}{}^{\nu} L - U^{\alpha\beta}{}_{,\mu} \frac{\partial L}{\partial U^{\alpha\beta}{}_{,\nu}} = U^{\alpha\beta}{}_{,\mu} U_{\alpha\beta}{}^{,\nu} - \frac{1}{2}\delta_{\mu}{}^{\nu} U^{\sigma\beta}{}_{,\rho} U_{\alpha\beta}{}^{,\rho} \tag{24}$$

Equations (22) and (23) lead to the field equations

$$\Box U^{\mu\nu} = 0 \tag{25}$$

(4) See, for example: Landau and Lifshitz, The Classical Theory of Fields, Addison Wesley.

(5) J. Weber, General Relativity and Gravitational Waves, Interscience Publishers, New York and London, 1961, page 73.

This does not include interaction of particles with themselves, according to (21a). We may obtain (21a) by adding a term f_1 to the Lagrangian density (22), obtaining

$$L' = -\frac{1}{2} U^{\mu\nu}{}_{,\alpha} U_{\mu\nu}{}^{,\alpha} + f_1 \tag{26}$$

The addition of f_1 to (26) leads to a new expression for $t_\mu{}^\nu$, which we denote by $t_\mu{}^{\nu\prime}$: so we require

$$\Box U^{\mu\nu} = t'^{\mu\nu} \tag{21b}$$

In order to obtain (21b) we must add a term to (26). But this leads to $t''_\mu{}^\nu$ which in turn leads to the requirement that new field equations

$$\Box U^{\mu\nu} = t''_\mu{}^\nu \tag{21c}$$

be obtained. By continuing this process a Lagrangian density with an infinite number of terms is obtained. Gupta[6] carried these procedures through and found that this Lagrangian density with an infinite number of terms is indeed equal to the curvature scalar, the correct one required to deduce the equations of general relativity.

To summarize, the equations of general relativity can be deduced by using the same philosophical notions as in other field theories. We deal with particles with spin 2, and

(6) S. N. Gupta, Phys. Rev. 46, 1683 (1954)

recognize that energy is the source of the gravitational field. The gravitational field itself contributes to part of the energy density, which in turn is a source of gravitational fields. This, then, leads to a non-linear theory, to a Lagrangian density L with an infinite number of terms, from which the curvature scalar can be obtained.

6. Electrodynamics in Arbitrary Coordinates and its Geometrization

Equations (14) and (15) are the four potential formulation of electrodynamics, in Lorentz frames. The field tensors $F_{\mu\nu}$ are given by

$$F_{\mu\nu} = A_{\nu,\mu} - A_{\mu,\nu} \tag{27}$$

and the Maxwell equations[7] for $F_{\mu\nu}$ are

$$F_{\mu}{}^{\nu}{}_{,\nu} = j^{\mu} \tag{28}$$

$$\frac{\partial F_{\alpha\beta}}{\partial x^{\gamma}} + \frac{\partial F_{\gamma\alpha}}{\partial x^{\beta}} + \frac{\partial F_{\beta\gamma}}{\partial x^{\alpha}} = 0 \tag{29}$$

For any $F_{\mu\nu}$ defined in terms of a four potential by (27), (29) is satisfied as an identity in consequence of the vanishing of $\nabla\cdot\nabla\times\bar{A}$. In arbitrary coordinates (28) becomes

$$F^{\mu\nu}{}_{;\nu} = j^{\mu} \tag{30}$$

(7) In this section we use the metric tensor $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$ to raise and lower indices.

We are using a semicolon to denote covariant differentiation.

Equation (29) becomes

$$\left(\epsilon^{\alpha\beta\gamma\delta}(-g)^{-1/2}F_{\alpha\beta}\right)_{;\delta} = 0 \tag{31}$$

In (31) $\epsilon^{\alpha\beta\gamma\delta}$ is the Levi Civita tensor density, $\epsilon^{0123} = 1$, it changes sign on interchange of any pair of indices and vanishes if two or more indices are the same. g is the determinant of $g_{\mu\nu}$. Using the standard formula for the covariant divergence of an antisymmetric tensor then gives us again (29) as valid in arbitrary coordinates. Using the standard formulas for covariant differentiation then leads to

$$A^{\nu,\mu} - A^{\mu,\nu} = A^{\nu;\mu} - A^{\mu;\nu} = F^{\mu\nu} \tag{32}$$

Making use of the generalized Lorentz gauge condition

$$A^{\mu}{}_{;\mu} = 0 \tag{33}$$

and the rules for changing the order of covariant differentiation then reduces (30) to

$$A_{\mu;\alpha}{}^{;\alpha} - R_{\mu\alpha}A^{\alpha} = -j_{\mu} \tag{34}$$

Here $R_{\mu\nu}$ is again the Ricci tensor. It seems from this that electrodynamics fits very naturally into the scheme of general relativity. But Einstein felt that this was not enough. Since the geometrical interpretation of gravitation was successful, he thought perhaps one ought to try to geometrize electromagnetism.

Part of the motivation for the geometrization is the fact that in general relativity the gravitational forces are entirely taken into account by the geometry of the space. With electromagnetic forces, the spinless particles no longer move along geodesics, if they are charged.

For a long time it was believed that a partial geometrization of gravitation and charge free electromagnetism could be achieved by elimination of the Maxwell field tensor[8], [9] from the coupled Maxwell Einstein equations. This can be accomplished in consequence of some quite special properties of the Maxwell tensor. These are

$$T_\alpha{}^\alpha = 0 \tag{35}$$

$$T_\mu{}^\alpha T_\alpha{}^\nu = \frac{1}{4}\,\delta_\mu{}^\nu T_{\alpha\beta}T^{\alpha\beta} \tag{36}$$

$$T_{oo} > 0 \tag{37}$$

These relations lead to the following equations

$$R = 0 \tag{38}$$

$$R_\alpha{}^\beta R_\beta{}^\alpha = \frac{1}{4}\,\delta_\alpha{}^\gamma R_{\sigma\tau}R^{\sigma\tau} \tag{39}$$

(8) G. Y. Rainich, trans. Am. Math. Soc. 27, 106 (1927).

(9) C. W. Misner and J. A. Wheeler, Annals of Physics 2, 525, (1957)

$$\left[\frac{\epsilon_{\beta\lambda\mu\nu}R^{\lambda\gamma;\mu}R_{\gamma}{}^{\nu}\sqrt{-g}}{R^{\sigma\tau}R_{\sigma\tau}}\right]_{,\alpha} = \left[\frac{\epsilon_{\alpha\lambda\mu\nu}R^{\lambda\gamma;\mu}R_{\gamma}{}^{\nu}\sqrt{-g}}{R_{\sigma\tau}R^{\sigma\tau}}\right]_{,\beta} \tag{40}$$

The metric which satisfies these relations has as its source the stress energy of a field satisfying Maxwell's equation. It was shown by Witten[(10)] and independently by Penrose that the required Cauchy data to integrate these equations may correspond to more than one Maxwell field. This description is therefore not unique. The Maxwell tensor cannot be eliminated without elimination of at least part of the physics.

The attempts to achieve a complete geometrization[(11)] have extended over many years. When this program was begun it was believed that gravitation and electromagnetism comprised all of physics. Now geometrization would have to include quantum effects as well as the strong and weak interactions. The extraordinary difficulty of such a program has resulted in its abandonment by all but a very few mathematicians and physicists.

---

(10) L. Witten, Phys. Rev., 120, 635, (1960)

(11) V. Hlavaty, <u>Geometry of Einstein's Unified Field Theory</u>, P. Noordhoff Groningen, Netherlands, 1957.

7. Quantization of the Coupled Maxwell Einstein Fields

Electrodynamics and Gravitation may be written in Hamiltonian form and a quantization carried out in an approximation scheme[12]. For weak fields we write

$$g_{\mu\nu} = \delta_{\mu\nu} + h_{\mu\nu} \tag{41}$$

We will use Latin letters for the space indices 1, 2 and 3. Coordinates may be chosen such that

$$g_{\mu o} = \delta_{\mu o} \tag{42}$$

With these assumptions the Hamiltonian for the coupled Maxwell Einstein fields may be written in the approximate form

$$H = H_G + H_M + \int (h_{rs}\eta_r\eta_s/\rho - 2h_{\ell j}F_{m\ell}F_{dj}\delta^{md})\, d^3x \tag{43}$$

In (43) H_G contains only the gravitational field variables and momenta, H_M contains only the Maxwell field variables A_k and the canonical momenta η_k. Let us consider the interaction terms in (43), when the theory is quantized. It is seen to be made up of sums of products, each containing one gravitational field operator, and two Maxwell field operators. This interaction implies that a photon

---

(12) See, for example, J. Weber and G. Hinds, Physical Review, to be published.

can decay into another photon and a graviton (Figure 3)

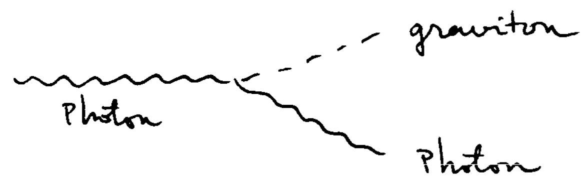


Figure 3

A careful study of this process shows that the matrix elements for it do not vanish unless all three particles propagate in the same direction. However, all three particles have zero rest mass. Energy and momentum can be strictly conserved only if all particles propagate in the same direction. Therefore, this process cannot occur except conceivably at extreme energy where strict conservation of energy can be somewhat relaxed. By extreme energies we mean energies >> 10^{28} electron volts. This kind of process cannot therefore explain the red shift as a "tired light" mechanism during the long propagation time from distant galaxies.

Further study of the interaction shows that we may expect graviton production if photons are incident on a coulomb field or a magnetostatic field. The cross section is very small. For a coulomb scatterer containing uniform electric or magnetic fields, with linear dimensions all large compared with the wavelength of the incident photon the cross section for this process is

$$S = \frac{8\pi^2 GU\ell}{c^4} \tag{44}$$

Here U is the energy of the scatterer, and ℓ its linear dimension in the direction of propagation of the photon. For laboratory experiments the cross section appears much too small. Thus, a cubic meter containing 10^{15} ergs of electrical energy has $S \approx 10^{-30}$ cm^2. A galaxy with a magnetic field $\approx 10^{-6}$ gauss would have a cross section $\approx 10^{28}$ cm^2 and convert roughly one part in 10^{16} of the incident photons to gravitons by this process. We note the absence of Planck's constant in (44). The interaction of two Boson fields has a classical limit and this is expressed by (44).

POSSIBLE EFFECTS ON THE SOLAR SYSTEM OF φ WAVES IF THEY EXIST

Lecture XII

R. H. Dicke

Princeton University

Seminar on Gravitation and Relativity, NASA GOddard Space Flight Center, Institute for Space Studies, New York, N. Y.; edited by H. Y. Chiu and W. F. Hoffmann.

Preface

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R. H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution the space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a minimum amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and correct the final manuscript. Hence, we accept responsibility for errors and omissions.

H. Y. Chiu

W. F. Hoffmann

A φ wave is a wave in a long-range scalar field. The question of the existence of these waves is probably the most important part of the title of this lecture and the part about which we can say the least. On the other hand we can describe with reasonable confidence the basic properties of these waves which follow from the requirement of relativistic invariance and the results of certain experimental observations. So we are in the unusual situation of knowing more about the properties of this field than its existence.

I shall briefly review some of these properties which were discussed in Lectures VII and VIII. In those two lectures a long range scalar field was introduced into gravitation theory in order to modify general relativity in such a way as to make it more compatible with requirements of Mach's principle. In such a modified theory there are two alternative mathematical forms which the equations may take. In the first form matter behaves in an ordinary fashion; that is, the rest mass and physical dimensions are constant from place to place. But in this form the Einstein field equations are not valid. In the second form of the theory the Einstein field equations are valid but the scalar field, instead of appearing as part of the description of the description of the gravitation, appears as an ordinary matter long-range interaction which gives rise to non-constant particle dimensions. For the first form, the Jordan-type theory, the variational principle has the form:

$$0 = \delta \int \left[\varphi R + \frac{16\pi}{c^4} L - \frac{\omega \varphi_{,i} \varphi^{,i}}{\varphi} \right] \sqrt{-g} \, d^4x \tag{1}$$

The first term, from which one obtains the field equations for the components of the metric tensor, contains the scalar curvature, R, multiplied by the scalar field, φ. The presence of φ in this term is responsible for the departure of these equations from the Einstein form. The second term, involving only the matter Lagrangian L, yields the usual geodesic equations for the motion of particles. The last term gives rise to a wave equation for the scalar field, φ:

$$\Box \varphi = \frac{8\pi}{(2\omega + 3)} T \tag{2}$$

$\Box\varphi$ is the d'Alemtertian of φ, T is the contracted energy momentum tensor for all particles and fields, and ω is a dimensionless coupling constant for the scalar field. ω is of the order of magnitude of unity. A comparison with observations suggests a value of approximately 6. The gravitational constant, G, which does not appear explicitly in the variation principle is determined by the value of φ $(G \sim \frac{1}{\varphi})$.

The second form of the theory can be obtained by a transformation which corresponds to a redefinition of the units. Under this transformation the unit of length is changed by a scale factor which is a function of φ and the variational principle takes the form:

$$0 = \delta \int \left[\bar{R} + \frac{16\pi G}{c^4} (\bar{L} + \bar{L}_\varphi) \right] \sqrt{-g} \, d^4x \qquad (3)$$

The new curvature scalar, $\bar{R}$, is obtained by a conformal transformation on the old one. The new Lagrangian density for matter, $\bar{L}$, is modified due to the transformation of units. $\bar{L}_\varphi$ is the new Lagrangian density of the scalar "matter" field. In this form of the theory, the Einstein field equations for the components of the metric tensor are valid but the equations of motion of particles are modified. The scalar field enters as a long range interaction of matter rather than as part of the description of the gravitation field (hence the geometry). Physically, both forms of the theory are equivalent. However, for the purpose of discussing φ waves, the second form given by Equation (3) is more convenient. It is easier to visualize the effects of an ordinary long range matter interaction than the effects of the complicated coupling between the φ field and the metric tensor as given by Equation (1). The wave equation for φ from Equation (3) is:

$$\Box(\ln\varphi) = \frac{8\pi G T}{c^4(3+2\omega)} \tag{4}$$

In this form of the theory the mass of a particle is a function of the variable, φ, such that:

$$m = m_0 \varphi^{-1/2} \tag{5}$$

where m_0 is a constant. We may introduce this into Equation (4) by writing the contracted energy momentum tensor of matter:

$$T = T_0 \varphi^{-1/2} \tag{6}$$

where T_0 does not contain φ explicitly. Then Equation (4) becomes:

$$\Box(\ln\varphi) = \frac{8\pi G}{c^4(3+2\omega)} T_0 \varphi^{-1/2} \tag{7}$$

An interesting property of this equation is that for matter in a localized bound system, occupying a certain fixed volume over a long time average, the virial theorem implies that $T_0\varphi^{-\frac{1}{2}}$ is the integral over the volume of the total energy of this system. However, $T_0\varphi^{-\frac{1}{2}}$ is not a strictly conserved quantity in this theory. For example, for a radially oscillating star the integral of this quantity over the star is an oscillating function of time. It oscillates about a mean value which represents the total energy of the star (Figure 1).

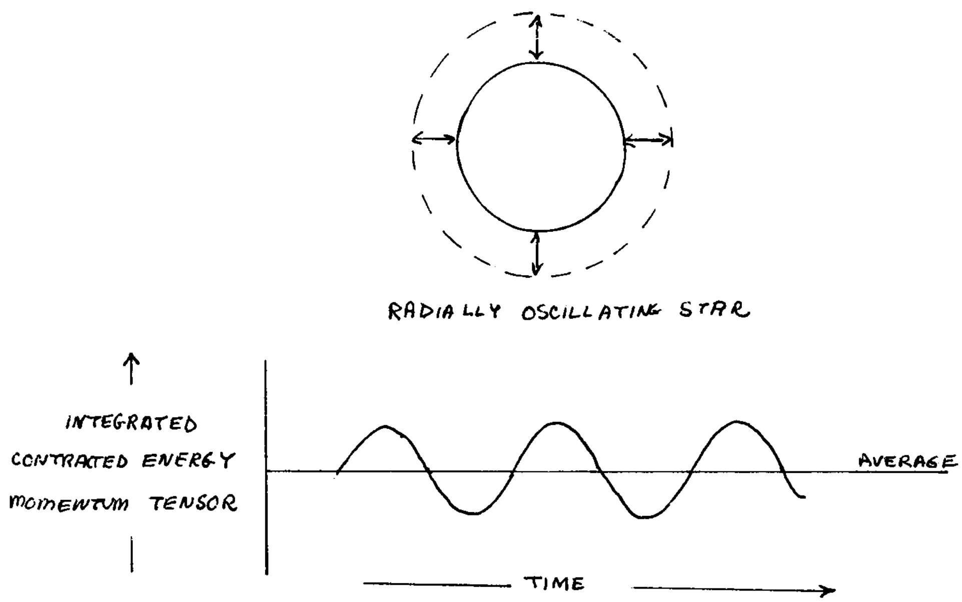


Figure 1. Radially Oscillating Star Has an Oscillating Contracted Energy-Momentum Tensor.

Such an oscillating star provides an oscillating monopole source for radiation of φ waves which, in principle, can be detected elsewhere in space. This would be a new phenomenon which should not occur for the ordinary gravitational field. It is a well-known property of general relativity that a localized radially oscillating star radiates no gravitational waves. This is because gravitational radiation, according to general relativity, is polarized quadrupole radiation. A radially oscillating star has spherical symmetry and cannot produce such polarized quadrupole radiation. We shall return to this in connection with the case of a collapsing star.

Another phenomenon associated with scalar monopole gravitational waves has been studied by Dieter Brill. He considered the radiation from planets moving in elliptical orbits. It is conceivable that monopole scalar field radiation would be so strong as to provide a damping mechanism for planetary motion that would be incompatible with observations. A planet moving in an eccentric elliptical orbit contributes to an oscillating contracted energy momentum tensor integrated over the solar system (Figure 2).

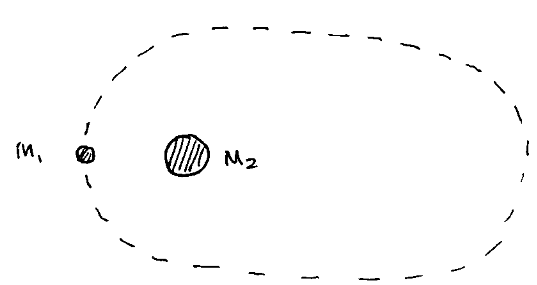


Figure 2. Monopole Radiation for a Two Particle Gravitational System

Hence monopole waves should be radiated. However, it turns out that this radiation is as weak as conventional general relativistic radiation due to the oscillating quadrupole moment of the solar system. The reason for this is that the gravitation quadrupole moment, which depends on the unsymmetrical distribution of the total mass of the system, oscillates through its full value. On the other hand the oscillating part of the monopole source depends on the relatively small variation in the kinetic energy of the system due to the elliptical planetary motion. This kinetic energy is of the order of $(v/c)^2$ times the total energy, where v is the velocity of the planets: $(v/c)^2 \approx 10^{-8}$. The quadrupole radiation rate is proportional to the quadrupole moment times $(v/c)^4$. The monopole radiation rate is proportional to the monopole moment times $(v/c)^2$. Since, from the above argument, the monopole moment is roughly $(v/c)^2$ times the quadrupole moment, the radiation rate of a monopole is roughly the same as that of a quadrupole.

A collapsing star might provide a much stronger source of φ waves. The core of a star may collapse due to rapid thermodynamic change of state which occurs at the end of thermonuclear evolution. At present it is understood that supernova explosions are triggered by such a collapse.

During the collapse phase the quasi-state equilibrium state of the core is destroyed suddenly and the material of the star falls freely inward. The contracted mass energy tensor $T_0 \varphi^{-\frac{1}{2}}$ may change by as much as one hundred percent. The energy of the star which could be radiated as a φ wave during such a collapse is in order of magnitude:

$$\frac{\text{ENERGY IN } \varphi \text{ WAVE}}{\text{TOTAL ENERGY OF STAR}} \sim \frac{1}{(3+2\omega)} \left(\frac{GM}{Rc^2} \right)^{3/2} \qquad (8)$$

where M is the mass of the star and R is the radius from which it starts to collapse. This radius can be very small since as the core approaches the critical mass, it shrinks and becomes a degenerate star of a very small radius. (For a fuller discussion, see Lecture X, "Degenerate Stars" by J. A. Wheeler.) Gm/Rc^2 may be of the order of 10^{-4}, or perhaps even larger. So we expect at least a millionth of the energy, perhaps much more, to be radiated in a wave of this kind.

The fraction of energy radiated depends on the time scale of the rapid collapse. An alternate way of writing Equation (8) in terms of this time of collapse is:

$$\frac{E_{\varphi\text{-WAVE}}}{E_{\text{STAR}}} \sim \frac{1}{(3+2\omega)} \frac{GM}{Tc^3} \qquad (9)$$

where T is on the order of a few seconds.

It is very difficult to find mechanisms in the universe as it exists now for radiating any larger amounts of this energy. Therefore, it would be difficult to believe that there are sources for this kind of a field which would lead to an energy density in space which is comparable to that of ordinary matter.

On the other hand if the universe has evolved from a highly compressed state, it is possible that the early evolution of the universe could have generated a large density of φ waves that could have persisted until now. This could lead to a substantial part of the energy density of space being in the form of scalar field waves.

The average energy density in scalar field waves is related to the rate of change of the field in the following way:

$$\bar{u} \sim \frac{(3+2\omega)}{16\pi G} c^2 \overline{\left(\frac{\dot{\varphi}}{\varphi}\right)^2} \tag{10}$$

$\dot{\varphi}$ is the time derivative of the field φ.

We can estimate the maximum possible value for $(\dot{\varphi}/\varphi)$ by assuming the energy density of these waves to be the average energy density required by the cosmological model for the universe expanding at a rate given by the Hubble

Constant. This energy density is about $10^{-29} g/cm^3$. Substituting this into Equation (10) gives:

$$\sqrt{\overline{\left(\frac{\dot{\varphi}}{\varphi}\right)^2}} \simeq \frac{1}{2 \times 10^{10} \text{ years}} \qquad (11)$$

This is an average fractional change in the field of order of magnitude of one part in 10^{10} per year. Hence, if this energy density is relatively uniformly distributed, its effects would be very small and extremely difficult to observe. However, it is not clear that it is necessary to assume uniform distribution.

The reason for this is that the non-linear character of the wave equation (Equation (7)) provides some mechanisms that tend to sharpen wave fronts. This can be seen qualitatively in the following way.

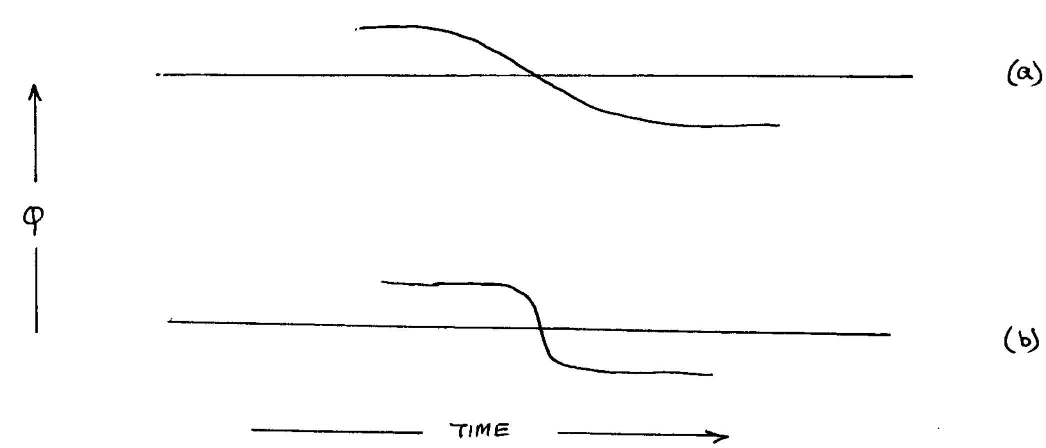


Figure 3. The Sharpening of a Wave Front as a Wave Passes Through Matter.

Figure 3(a) shows φ at a point in space as a wave front. As this wave passes through matter so that the local value of φ is increased, the energy associated with the matter is decreased as a consequence of the dependence of the mass of particles on φ (Equation (5)). Since the total energy is conserved, energy is transferred from the matter into the φ wave front. A more detailed analysis suggests that in this process the wave front can be sharpened.

Another interesting effect of a φ wave passing through matter is the φ-wave maser. This might operate on the scale of a galaxy. In a galaxy of some 10^{10} hydrogen burning stars many of these are in the white dwarf stage for which the core size is approaching the critical Chandrasekhar limit beyond which they will undergo unstable collapse. A slight increase in the gravitational constant would lower the critical mass so that all those stars that were nearing the collapse point would suddenly have exceeded it and as a consequence begin an unstable collapse. This collapsing star then radiates a φ wave contributing to the strength of the wave front, as previously described. Thus we can visualize a φ wave passing through a galaxy initiating the collapse of many white dwarf stars which then contribute wavelets to maintaining and strengthening the wave front.

Conceivably this model could be used to explain associated production of supernova which has been postulated as an explanation for the very intense observed radio sources. However the necessary calculations have not been made. These sources appear to be radiating energy at a rate too great to be due to a single supernova. It is necessary to assume a large number of supernova to produce radio sources this strong. Burbidge has argued that in the center of the galaxy there are many stars which have reached a critical state, about ready to become supernovae. By chance one goes off and produces a shock-wave that sets others off. Unfortunately, it has never been made quite clear how one supernova would set off another.

The φ wave model provides a possible mechanism for associated production of supernova. A galaxy with 10^5 or 10^6 stars about ready to explode encounters a φ wave in the form of an extraordinary large bump in the gravitational constant. All those stars that are ready to go, go all at once. This stirs up the gas sufficiently to provide a very strong radio source.

There is another rather interesting effect which can occur to a degenerate star which has a mass very near the critical mass. Under these conditions its equilibrium radius and energy are a very sensitive function of the gravitational constant. A change in perhaps one part in 10^6 of the gravitational constant could affect the total energy to the order of one percent or so. If such a greatly contracted star were intercepted by a φ wave corresponding to a weakening of gravity the star would puff up to reach its new equilibrium size, and to do so would absorb some energy from the φ field. On the other hand, if the φ wave corresponded to an increasing gravitational constant, then it would cause a further contraction and decrease in energy of the star. In this way the equilibrium energy of the star could be a rather sensitive function of the φ field. If the wave front were sufficiently sharp so that the star could not follow the change in φ quasi-statically through a series of equilibrium states, the star might pulsate for a while about the new equilibrium state after the wave front has passed. This could lead to φ wave radiation.

We have been speaking mainly of the effects of single φ wave fronts without being concerned with what frequency range would characterize these waves. It is difficult to say what this range should be. The mechanisms for production of φ waves are not likely to be on the atomic scale. The coupling strength of this field is of the order of 10^{-40} of other atomic coupling strengths. Hence competing radiation and energy exchange processes would rule out significant φ wave radiation on an atomic scale.

More likely mechanisms for φ wave radiation involve coupled phenomena where many particles move together. Systems of particles as large as the sun do not move together unless they move relatively slowly. Thus a reasonable lower limit for the period of φ waves is given by the free fall time of a degenerate star. This time is in the range of seconds. However, if the main source of φ wave radiation occurred at the time the expansion of the universe started, this radiation would have been red shifted ever since and periods of seconds might have been shifted into hours or days. Other processes could have produced waves with periods now of the order of 10's of years. We really do not know the period to expect for φ waves except that the range of hours to tens of years might be in a reasonable range.

It appears that if such fields exist they could have some interesting effects on galaxies and stars. Another interesting question is whether a field of this kind would have effects on the solar system for which we have precision observations. To be able to say that the field exists on the basis of what one sees in the solar system is I think extremely unlikely. The earth and the planets are sufficiently complicated so that there are usually alternative explanations for the small effects we may observe. However, we can determine what the implications of φ waves impinging on the solar system would be, and then ask whether we can rule out a field of this kind on the basis of what we see.

As a starting point I am going to assume the largest rate of G variation that might have escaped detection by present methods of observation. This rate corresponds to a fractional change in one year:

$$\frac{\Delta G}{G} \simeq 10^{-8} \tag{12}$$

This variation is considerably larger than that due to the secular rate of change of φ of the order of 3 parts in 10^{11} associated with the cosmological solution of the scalar theory for an expanding closed universe (Lecture VIII). In fact, if space were filled with a φ field whose time rate of change corresponds to this variation in G, the energy density of the field would be 10^4 times that permissible on astronomical grounds (Equation (9)). Hence on the average no more than 10^{-4} of space can be filled with φ waves of this strength. Thus the _a priori_ probability of such a wave impinging on the solar system in any given year is 10^{-4}. This is a small probability. Its major significance is that we cannot rule out the existence of φ waves on the basis of a lack of observation of the effects of a φ wave of this strength. Furthermore, if we cannot rule out a G variation this large by observation we certainly cannot rule out a more reasonable smaller variation.

What are the effects of strong φ waves on planetary orbits? For one thing, the eccentricity of a planetary orbit would be changed by the passage of a φ wave front (Figure 4). For example, if the φ wave causes a sudden decrease in G for some position (A) of the planet in a circular orbit, the planet will continue its motion in a new eccentric orbit, (B). The variation in eccentricity that would be expected from a change in G over a time short compared with the period of revolution of the planet is of the order of 10^{-8} with the above assumptions. We expect about 10^{-4} such disturbances per year. In the 4.5×10^9 years that the solar system has been in existence, there would have been of the order of 4.5×10^5 such jumps in the eccentricity. These should occur as a random walk. So that the expected departure of the eccentricity of a planet, initially in a nearly circular orbit, over 10^9 years would be

$$\Delta\epsilon \simeq \sqrt{4.5 \times 10^5} \times 10^{-8} \simeq 7 \times 10^{-6} \qquad (13)$$

This is much smaller than the planetary eccentricities which are observed so that this effect does not rule out the existence of φ waves. Nor have I found any effects which would have affected the forms of orbits in the solar system sufficiently in historical times to be ovservable.

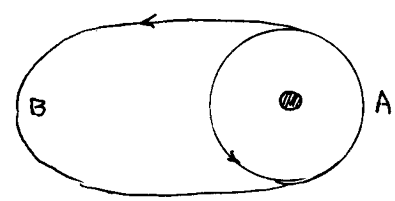


Figure 4. Effect on Planet Orbit of Abrupt Change in G.

On the other hand there are uncertainties, variations, or systematic discrepancies in planetary orbits that are not understood, and for which no one has an explanation. The fact that the orbits do not follow exactly what one predicts from conventional theory is a loophole for allowing disturbances of this kind to exist. For example, Clemence has pointed out that there seems to be a correlation in the residuals in Saturn's orbit and in Jupiter's orbit. It would be interesting to see whether these correlation effects could be tied to a common cause.

We may also ask about the effects of φ waves for which the gravitational constant changes over a period which is long compared to the orbit period. With such a slow (adiabatic) change, the orbit merely breathes in and out without changing its eccentricity. The most interesting place to look for an effect of this sort is with the moon's motion. If the gravitational constant were to have changed slowly by one part in 10^8 over 200 years, the moon would move now at a new rate compared with an atomic clock. Unfortunately we haven't had atomic clocks over the last 200 years. However, we can compare the moon's motion with the earth's rotation over this period of time.

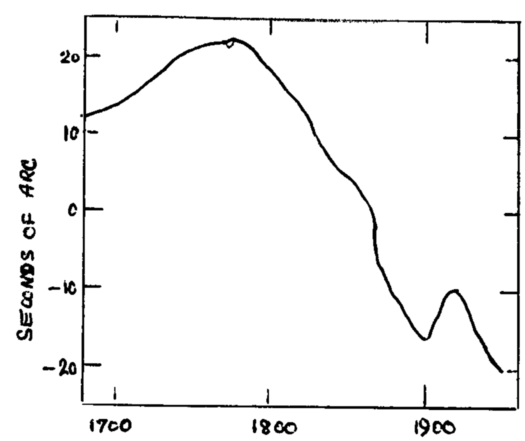


Figure 5. Discrepancy in the Moon's Longitude Relative to the Earth's Period

Figure 5 shows the lunar discrepancy curve over the last 200 years. The difference between observed and calculated positions are plotted in seconds of arc. The origin is arbitrarily chosen for a zero discrepance at 1870. In addition there is an arbitrary constant which establishes the initial ratio of the moon and earth 'clock' rates. It can be used to eliminate a linear term in the curve. The remaining wiggles and curvature are real and cannot be removed by an arbitrary choice of constants.

The points prior to 1900 fit quite well to a parabolic curve. This represents a uniform acceleration over this period with some wiggles superimposed on top of it. The really remarkable disturbance occurred rather recently in 1900, while fairly good observations of all kinds were being made. At this time there was a big bump in the discrepancy curve. One can interpret this particular bump in two different ways. One is to say that the rotation of the earth changed slightly. The other is to say that the period of the moon changed. I will describe the disturbance in terms of a change in the moon's period. That is, I will assume the moon started going around the earth at a different rate for a while and then it returned to its old rate.

Before doing this I will comment on whether the earth could have changed its rotation rate by a sufficient amount to account for this bump and how this could have happened. This question has been discussed rather completely by

MacDonald and Munk[1]. It is an historically old problem with which many astronomers have been very much concerned even before this very large discrepancy in 1900 was observed. To illustrate the extent of the problem I will read a statement made by the famous orbit astronomer, Newcomb, in 1909[2] and quoted by MacDonald and Munk:

"I regard these fluctuations as the most enigmatical phenomenon presented by stellar motions, being so difficult to account for by the action of any known cause that we cannot but suspect them to arise from some action in nature hitherto unknown."

Then MacDonald and Munk point out that "Sea level variations, continental unrest, melting on Antarctica and other observable processes cannot possibly be the cause. The only known hope is the core; we have arrived at this conclusion by what Sir Edward Bullard has called the Sherlock Holmes procedure, of eliminating one possibility after another."

---

(1) G. J. F. MacDonald and W. H. Munk, The Rotation of the Earth, Cambridge (1960).

(2) S. Newcomb, Fluctuations in the Moon's Motion, Monthly Notices, Royal Astronomical Society, 69, 164 (1909).

To account for such a large change in the earth's rotation rate requires unreasonably large disturbances on the earth. For example, the level of the sea would have to have changed by a meter to account for this change in the earth's rotation rate.

A further difficulty in reconciling this variation with known causes is the fact that the earth's pole did not change its location appreciably in this period of time. For example, the explanation in terms of a change in the ocean level of one meter due to melting of ice in Antarctica would have resulted in the North Pole of the earth moving some 100 to 200 feet from where it was. This is because the Antarctic ice is not distributed symmetrically about the earth's axis of rotation. The stability of the position of the earth's rotation axis over recent times requires that any mechanism postulated to cause a change in the moment of inertia of the earth by the necessary amount be so symmetrical as not to change the rotation axis.

Another possible cause for a change in the earth's rotation rate is a change in the angular momentum carried by the atmosphere. Unfortunately, this is roughly two orders of magnitude too small to account for the observed change in rotation rate.

Could it be due to continental blocks moving up and down? We know roughly how much continents are moving vertically from observations of sea levels variations. These motions are also orders of magnitude too small to produce the necessary effect on the earth's rotation rate. Furthermore, an explanation in terms of continental blocks moving would again run into trouble with the motion of the earth's pole.

Of all the effects that could occur near the earth's surface, nothing really fits. The only remaining possibility seems to be a possible change in the angular momentum of the earth's core. This is quite difficult to get at.

But there is some evidence that the magnetic field of the earth, which is presumed to be connected with currents in the core, has been drifting from east to west. Also, there is some indication that there was a change in the rate of drift of the magnetic field at the time of the large bump in Figure 5. Thus through some way not understood in detail there could have been a transfer of angular momentum to the earth's core to change its rotation rate.

Now I would like to turn to the possibility that the discrepancy is at least in part due to a variation in the moon's period. It is interesting that the change in the period in 1900 of four parts in 10^8 is even greater than the rate of change considered in the above discussion of the effects of φ waves. It would be difficult to exclude a variation of G accounting for at least a part of this discrepancy. It is even possible that a φ wave could account for all of the 1900-1920 disturbance.

This raises the interesting question: If this were actually a change in the moon's motion due to a change in G rather than a change in the earth's rotation, what other effects would be expected to be associated with this? We have concluded that one rather sensitive test for a variation of G is the frequency of earthquakes. The reason for this is the following:

Stresses across a fault plane in the earth build up slowly through normal tectonic processes. Lateral displacements of the order of one centimeter per year occur.

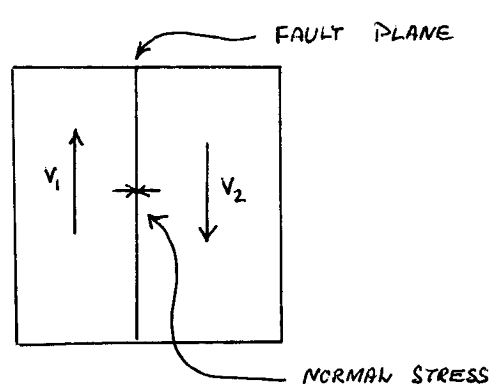


Figure 6. Illustration of the Inhibition of Earthquakes Along a Fault Plane Due to Normal Stresses.

The dimensions of a fault plane may be of the order of a thousand kilometers, with strains building up at the rate of one part in 10^8 per year. These strains continue to build up until the strength of the material is exceeded. Then the material flows or slips along the fault plane and earthquakes are produced.

It is clear from Figure 6 that the normal stresses across the fault plane will influence the ease with which motion along the fault plane can occur. Thus small changes in the normal stresses can provide a trigger mechanism for initiating or inhibiting earthquakes. If G were to become a little smaller, the earth would expand slightly and these normal stresses would get smaller regardless of the orientation of the fault plane. If G were to vary by the order of one part in 10^8 in a year, the resulting expansion of the earth would be of the order of a tenth of that, about one part in 10^9. However, the stresses along the fault planes are accumulating of the order of one part in 10^8 in a year. So the variation of normal stress is quite appreciable in comparison with the rate at which the stress is building up. Thus the determination of whether an earthquake should go or not can be rather strongly affected by a variation of G of that order of magnitude.

What effect then on earthquakes would we expect to be associated with the bump on the moon-earth rotation discrepancy curve in 1900? If at that time the moon started to move more rapidly, this implies that G became larger. This would make the normal stresses larger, having the effect of turning off earthquakes. About twenty years later G becomes smaller again; normal stresses become smaller and the earthquakes should occur more frequently. So we expect a period of 20 years with a low earthquake rate, followed by an enhanced earthquake rate during the period following a decrease in G.

The upper curve in Figure 7 shows the variation of the frequency of earthquakes over this period of time. Indeed, this data shows a very low earthquake frequency, followed by an enhanced frequency coinciding roughly with the bump in the curve in Figure 5. This coincidence encourages us to examine the comparison between the earthquake rate and the moon's motion in greater detail.

The lower curve in Figure 7 shows the year by year average of the moon's motion. The moon's motion is much more noisy in this curve than in Figure 5 because it has not been smoothed to the extent of the previous graph.

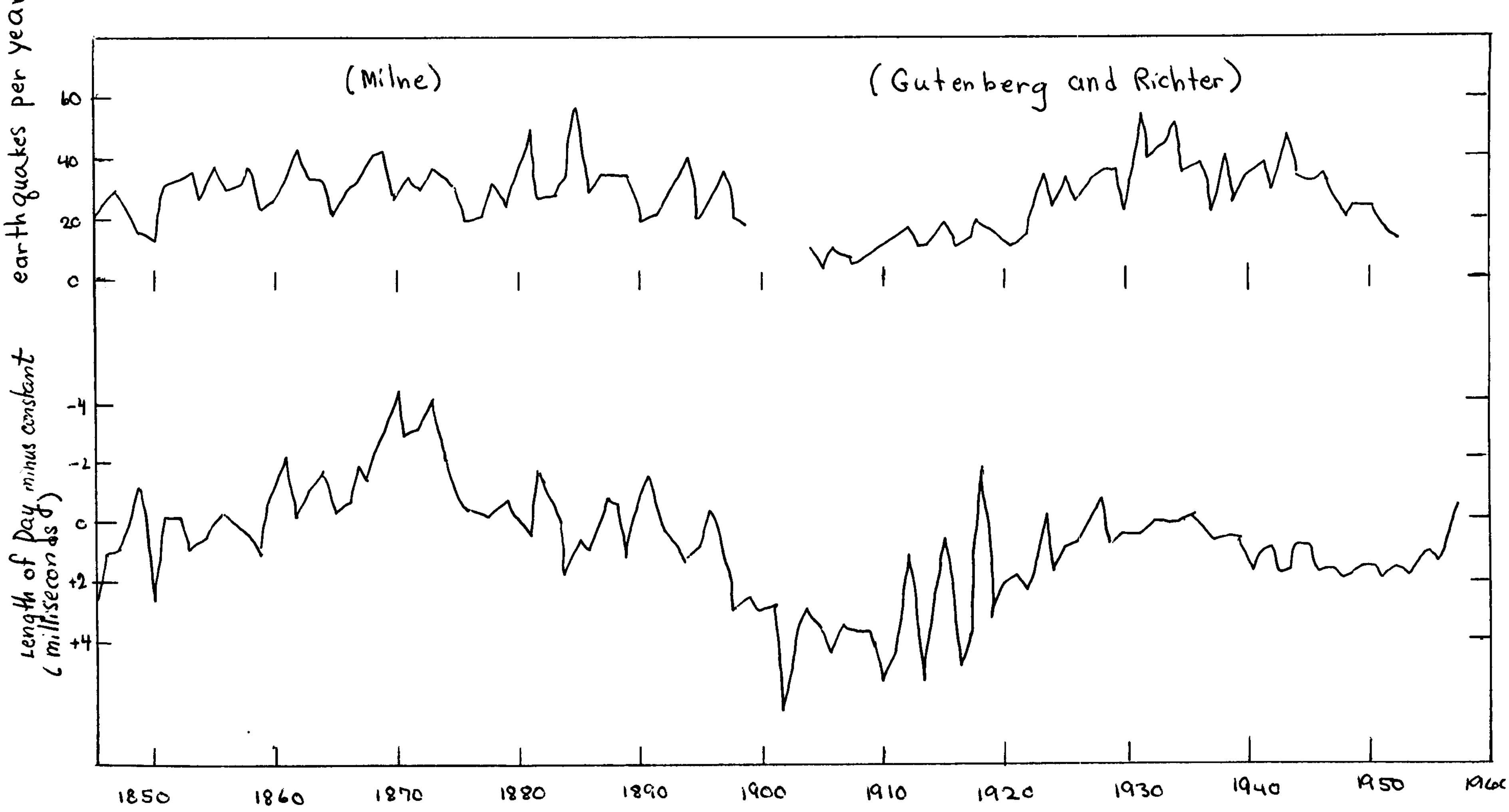


Figure 7. Comparison of number of earthquakes per year with the moon's motion. (The length of day was measured by comparing the rotation of the earth to the period of the moon, and so contains any information of a change in the moon's period.)

The data for the period of time from 1904 to 1952 was obtained from seismometers observations* and for the period prior to 1900 from newspaper reports. There is no immediately apparent correlation prior to 1900 between the newspaper reported earthquakes and the moon's motion. On the other hand I do not expect newspaper reports to be a source of information with a good signal to noise ratio. Rather I expect them to be rather poor.

One of my students, Jason Morgan, has calculated the correlation between earthquake frequency and the discrepancy in the moon's motion during this period. The correlation coefficients all have the right sign and have a reasonable significance level. He found the calculated correlation coefficient prior to 1900 is 0.26. The probability of obtaining a correlation that good by accident with random numbers is only of the order of three percent. The correlation for the period after 1900 during which the earthquake data is from seismometer observations is 0.71. The probability of getting this value accidentally with random numbers is very small, approximately 10^{-6}.

---

* All earthquakes of magnitude $M \geq 6.5$ is listed in B. Gutenberg and C. F. Richter, Seismicity of the Earth, 2nd ed., Princeton University Press, 1954.

Morgan examined in detail the data since 1900 to see where the correlation arises. Is it in only the low frequency fluctuations or are the high frequency variations meaningful? There are several very large bumps in the moon's motion on the lower curve in Figure 7. We might ask whether these big bumps correlate with bumps on the upper curve. It turns out that they do.

The analysis was made by fitting a cubic curve to these two plots and then subtracting the cubic off in order to leave only the "noise" fluctuations for which a correlation is computed. These two difference curves are shown in Figure 8. The cubic curves correlate very nicely with each other with a correlation coefficient of 0.94. The remainder correlates to the extent of 0.20. The probability of getting that kind of correlation with random numbers is eight percent. Thus all the correlations have the same sign and they are all reasonably above the level for which we would expect to get them through accident. These correlations fit very well with the hypothesis that (1) the disturbances on the earth-moon rotation rate discrepancy curve are associated with variations in the moon's period; (2) these variations and the variations in earthquake rates have a common origin in fluctuations of the value of the

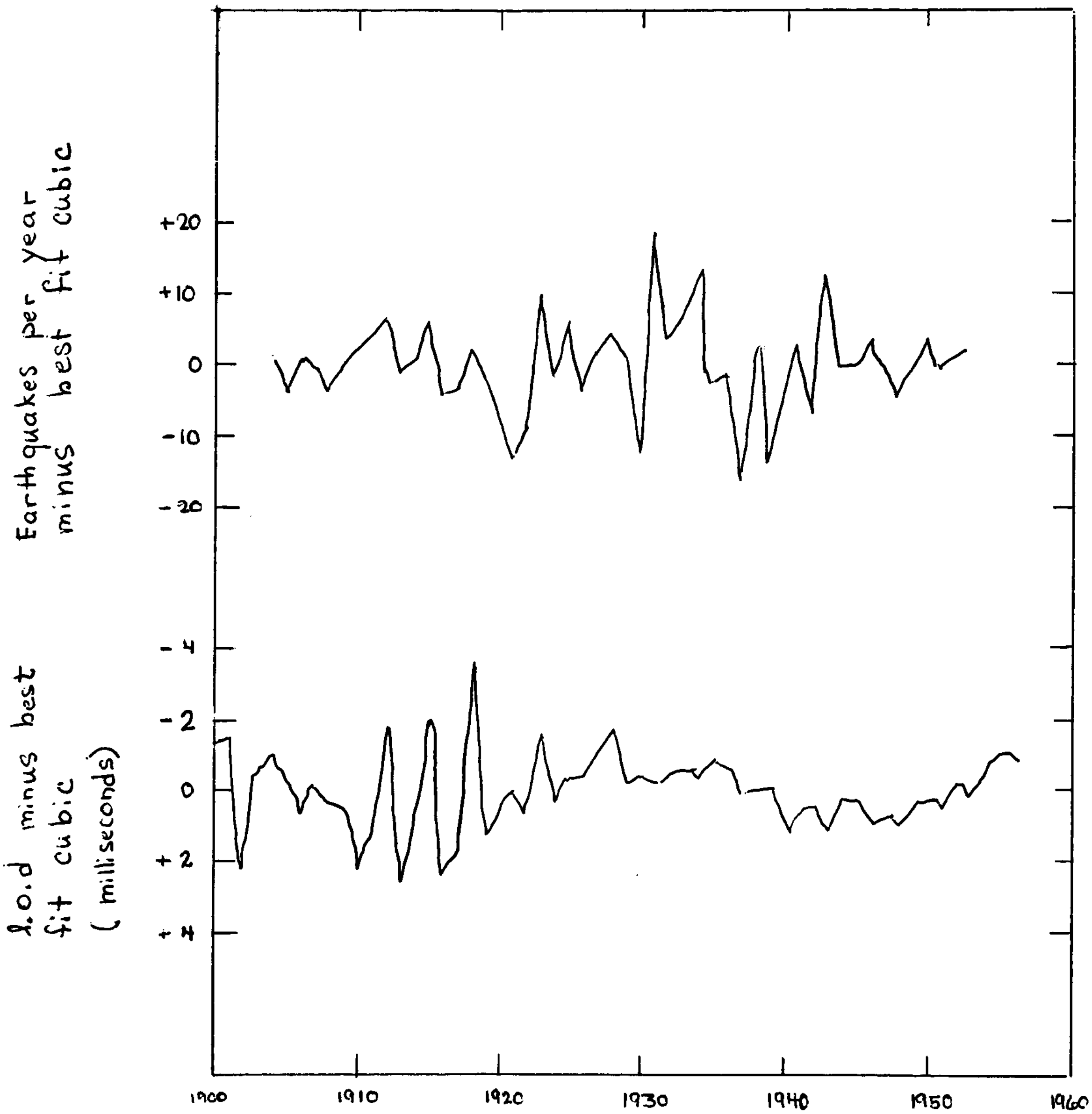


Figure 8. Length of day and earthquake frequency after subtracting cubic curves.

gravitational constant; and, (3) this could be caused by the passage of φ waves.

We may also ask the question whether the correlations fit the hypothesis of undisturbed motion of the moon and variations in the earth's rotation rate of some unknown origin. I believe that they can be made to fit but not as convincingly as they fit the hypothesis of irregular moon motion. If the earth changes its rotation rate in such a way as to speed up, there is a centrifugal expansion of the earth and a release of strains. The normal stresses become smaller by one percent of the fractional change in rotation rate (since centrifugal forces account for about one percent of g at the surface of the earth). If we wish to say that this is the mechanism for the strong correlation, then we must assume that earthquakes are extremely sensitive to small variations in normal stresses. However the fractional change in stresses produced in this way is only of the order of 10^{-10} which is about 10^{-3} of the tidal variation in stresses. Hence the tital effects would appear to dominate and mask any small effects of a varying earth rotation rate.

Another possible mechanism for the correlation in earthquake rate with the earth rotation changes is the variation in sheer stresses between the core and mantle. But if this were the case, the maximum earthquake rates should occur when the augular velocity of the earth is changing, rather than when the moon's angular velocity is a minimum as appears to be the case.

In the last few years there has been a very promising source of data on the subject of variation in earth rotation rate. This is the comparison of the earth's rotation rate with atomic clocks. As yet this comparison has not been made for a sufficiently long period of time to have any correspondingly good measure of the moon's motion. This is because the observation accuracy on the moon is so poor and its period so much longer than that of earth rotation that it is necessary to average data over a long period of time to achieve much accurately. So short term data on the moon's motion is extremely noisy and is not of much use in comparison to the very precise current measurement of the earth's rotation rate.

A change in the gravitational constant should have a small effect on the earth's rotation rate. As G decreases, the earth expands. The compressibility of the earth is such that the effect of a changing G on the earth's rotation rate ought to be about one tenth of the effect on the moon's motion. Morgan has looked at the correlation between earthquake rates and changes in the earth's rotation determined by comparison with atomic clocks over the past six years. He has found a correlation which has the same sign as previous correlations provided the effect is interpreted in terms of a changing G. But it has the opposite sign if the effect is interpreted in terms of a changing G. But it has the opposite sign if the effect is interpreted solely in terms of the earth's rotation.

While this and the previous effects discussed do not prove the existence of φ waves and the consequent G variations, they do provide a tantalizing invitation to explore further the hypothesis of scalar gravitational waves. They certainly do not give us grounds for rejecting the possibility of their existence.

THE LYTTLETON-BONDI UNIVERSE AND CHARGE EQUALITY

Lecture XIII

V. W. Hughes

Yale University

Seminar on Gravitation and Relativity, NASA Goddard Space Flight Center, Institute for Space Studies, New York, N. Y.; edited by H. Y. Chiu and W. F. Hoffmann.

Preface

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R. H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution the space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a minimum amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and correct the final manuscript. Hence, we accept responsibility for errors and omissions.

H. Y. Chiu

W. F. Hoffmann

1. Introduction.

Whether or not the electron charge magnitude exactly equals the proton charge magnitude is an interesting and fundamental question in physics. In this lecture I should like to discuss the theoretical arguments on this question, some implications to physics, astronomy and cosmology of a slight departure from charge equality, and the most recent experimental determinations of the electron-proton charge ratio.

As you know, experimental findings of the late 19th and early 20th centuries culminating in Millikan's oil drop experiment led to the conclusions that electric charges occur always as integral multiples of a smallest unit, and that the smallest unit for positive charge (the proton) is equal to the smallest unit for negative charge (the electron). Thus an atom or molecule which consists of equal numbers of electrons and protons should be electrically neutral. In 1932 the neutron was discovered and it was found to have zero charge. By now there are some 30 so-called elementary particles known, and each of these appears to have a charge of +1, 0, or -1 electron charge unit.

2. Implications of a Charge Difference.

Ideally elementary particle theory should predict the observed spectrum of the elementary particles including their charge and mass ratios. Modern quantized field theory can describe discrete particles but cannot predict the values of a particle's mass and charge. These must be obtained from experiment. The invariance of the theory under charge conjugation (the interchange of particle and antiparticle) does provide a theoretical prediction that a particle and its antiparticle should have charges which are equal in magnitude but opposite in sign. For example, the electron and positron charges should have the same magnitude. Also the proton and antiproton charges should have the same magnitude. However, theory does not predict the ratio of the magntidues of the charges on two different particles, for example, the ratio of the electron to proton charge.

Indeed in view of modern charge renormalization theory the question of the electron-proton charge ratio becomes rather deep and somewhat ambiguous. If the bare charges of the electron and proton were equal, then conventional renormalization theory with gauge invariance would require that the renormalized electron and proton charges should also be equal. However, Gell-Mann and Nambu [(1)] have remarked that if in

(1) M. Gell-Mann, Proceedings of the Tenth Annual International Conference on High Energy Physics (Interscience Publishers, Inc. New York (1960), p.792.

addition to the photon there were another neutral vector particle which is coupled to the proton but not to the electron, then even though the bare charges of the electron and the proton were equal, the renormalized charges would be expressed in terms of ambiguous, quadratically divergent integrals and might not be equal.

Feinberg and Goldhaber [(2)] have discussed the connection between the conservation laws and charge equalities of particles. At present the absolute conservation laws of charge, baryon number, and lepton number are all independent and are believed valid for any particle reaction. Because of the independent conservation laws for baryons and leptons, use of charge conservation in the known reactions involving elementary particles does not of itself determine the ratios of the charges of all the elementary particles. For example the apparent absence of the reaction $p \to e^+ + \pi^0$ leaves the ratio of the electron to proton charges undetermined. Conversely, if the electron (lepton) and proton (baryon) charge magnitudes were different, then the absence of such a reaction, or, more generally, the conservation of baryons would follow from the conservation

---

(2) G. Feinberg and M. Goldhaber, Proc. Natl. Acad. Sci. U. S. 45, 1301 (1959).

of charge instead of being an independent principle.

In the 20th century there has been considerable speculation about the effect on large-scale matter of a slight difference, δq, in the magnitudes of the electron and proton charges. Questions have been raised concerning the effect of such an inequality on gravitation, on the magnetic fields of astronomical bodies, and, recently, on cosmology.

As to the relevance of charge inequality to gravitation it is suggestive to compare the electrical force between two protons to their gravitational force. This ratio is:

$$\frac{F_{el.}}{F_{grav.}} = \frac{e^2/r^2}{Gm_p^2/r^2} = 1.2 \times 10^{36} \tag{1}$$

which is, of course, a very large number. If the electron charge is $q_e = -e$ and the proton charge were a slightly different magnitude,

$$q_p = (1 + y)e \tag{2}$$

then the charge on the hydrogen atom would be $+ye$, and the ratio of the electrostatic force between two hydrogen atoms to their gravitational force would be

$$\frac{F_{el.}}{F_{grav.}} = \frac{(ye)^2}{Gm_H^2} = 1.2 \times 10^{36}\ y^2 \qquad (3)$$

This ratio is 1 when $y = 0.9 \times 10^{-18}$. Hence if there were 1 part in 10^{18} difference between the proton and electron charge magnitudes, then the electrostatic force between two hydrogen atoms would be equal in magnitude to the gravitational force.

The very large ratio of electrical to gravitational forces and their similar dependence on the inverse square of the distance between the particles suggest the possibility that gravitational forces might arise due to some small breakdown of the normal theory of electrical forces. Lorentz proposed that the gravitational force might arise because of a slight difference between the force of repulsion between two particles with charges of the same sign and the force of attraction between two particles with charges of the same magnitudes but of unlike sign. Swann [(3)] has also discussed this possibility and has considered it in connection with matter and antimatter.

---

(3) W. F. G. Swann, Phil. Mag. 3, 1088 (1927); Astrophysical J. 133, 733 (1961).

The origin of the magnetic fields of astronomical bodies is another problem for which the possibility of a slight charge difference may be relevant. Einstein [4] remarked that a slight difference between the proton and electron charge magnitudes would, of course, lead to a net volume charge for matter composed of equal numbers of protons and electrons. Hence a rotating object such as the earth would have an associated magnetic field similar to that of a magnetic dipole. At the pole the field would be given by:

$$H_{pole} = 2P/R^3 \tag{4}$$

where R is the radius of the earth and P is its magnetic dipole moment:

$$P = \frac{0.2\omega MR^2}{c} \frac{\sigma}{\rho} \tag{5}$$

where ω is the angular velocity of the earth, M is the mass, σ is the charge density and ρ the mass density. For a proton charge given by equation (2)

$$\frac{\sigma}{\rho} = \frac{ye}{m_H} \tag{6}$$

where m_H is the mass of the hydrogen atom. If we assume that the earth's magnetic field of 0.6 gauss at the pole is entirely due to this charge inequality, then $y = 3 \times 10^{-19}$.

Blackett [5] observed in 1947 that the ratios of the magnetic dipole moment as computed from equation (5) to the angular momentum for three astronomical bodies--the earth, the sun and the star 78 Virginis--have nearly the same value of

$$\frac{P}{I}_{\text{earth, sun, star}} \simeq 1.1 \times 10^{-15} \tag{7}$$

Furthermore, the ratio of the orbital magnetic moment to the orbital angular momentum for an electron is

$$\frac{P}{I} \text{ electron orbital motion} = \frac{e}{2m_e c} \cong 0.9 \times 10^{7} \tag{8}$$

and the ratio of these two quantities is

$$\frac{(P/I) \text{ astronomical bodies}}{(P/I) \text{ electron}} \simeq 10^{-22} \tag{9}$$

This dimensionless ratio is nearly equal to the dimensionless constant

$$\frac{G^{\frac{1}{2}} m_e}{e} = 4 \times 10^{-22} \tag{10}$$

Blackett considered it unlikely that this approximate numerical equality should occur accidentally. Therefore he proposed that it should be true in general that

---

(5) P. M. S. Blackett, Nature 159, 658 (1947).

$$(P/I)_{\text{astronomical body}} = (P/I)_{\text{electron}} \frac{G^{\frac{1}{2}} m_e}{e} = \frac{G^{\frac{1}{2}}}{2c} \qquad (11)$$

It was found subsequent to Blackett's paper that the magnetic field of the sun is nearer to 1 gauss than to 50 gauss which was the value he used, so the ratio P/I for the sun actually does not have the value given in equation (7). There are many more stars whose magnetic fields have been determined by now and it would be interesting to compare these new data with equation (11).

The relation (11) is consistent with the model of a rotating charged earth that Einstein proposed. However, the simplest model of a rotating charged body gives very much too high an electric field at the surface of the earth so that the theory must be modified to include surface charge as well as volume charge in order to give a reasonable value for the electric field as well as for the magnetic field.

A third general area in which an electron-proton charge inequality might have some interesting implications is cosmology. Lyttleton and Bondi [6] suggested that the observed expansion of the universe might be understood in terms of a slight charge difference as an electric repulsion.

---

(6) R. A. Lyttleton and H. Bondi, Proc. Roy. Soc. A 252, 313 (1959).

They discussed this suggestion first in the context of simple Newtonian theory using the model of a smoothed out, spherical universe composed of hydrogen atoms with a mass density σ and a corresponding charge density σ, where

$$\sigma = \rho \frac{ye}{m_H} \tag{6}$$

and y is assumed to be positive. (See Figure 1.) The electrostatic force on a hydrogen atom at a distance r from the center of this charge distribution is

$$F_{el.} = \frac{(ye)^2}{r^2 \; m_H} \; M_r \tag{12}$$

where M_r is the total mass within the radius r.

The gravitational force is

$$F_{grav.} = \frac{M_r \; m_H G}{r^2} \tag{13}$$

We define the ratio of the electrostatic repulsive force to the gravitational attractive force to be

$$\mu = \left(\frac{ye}{m_H \quad \sqrt{G}} \right)^2 = (1.12 \times 10^{18} \; y)^2 \tag{14}$$

which is the same as equation (3). The net repulsive force is then

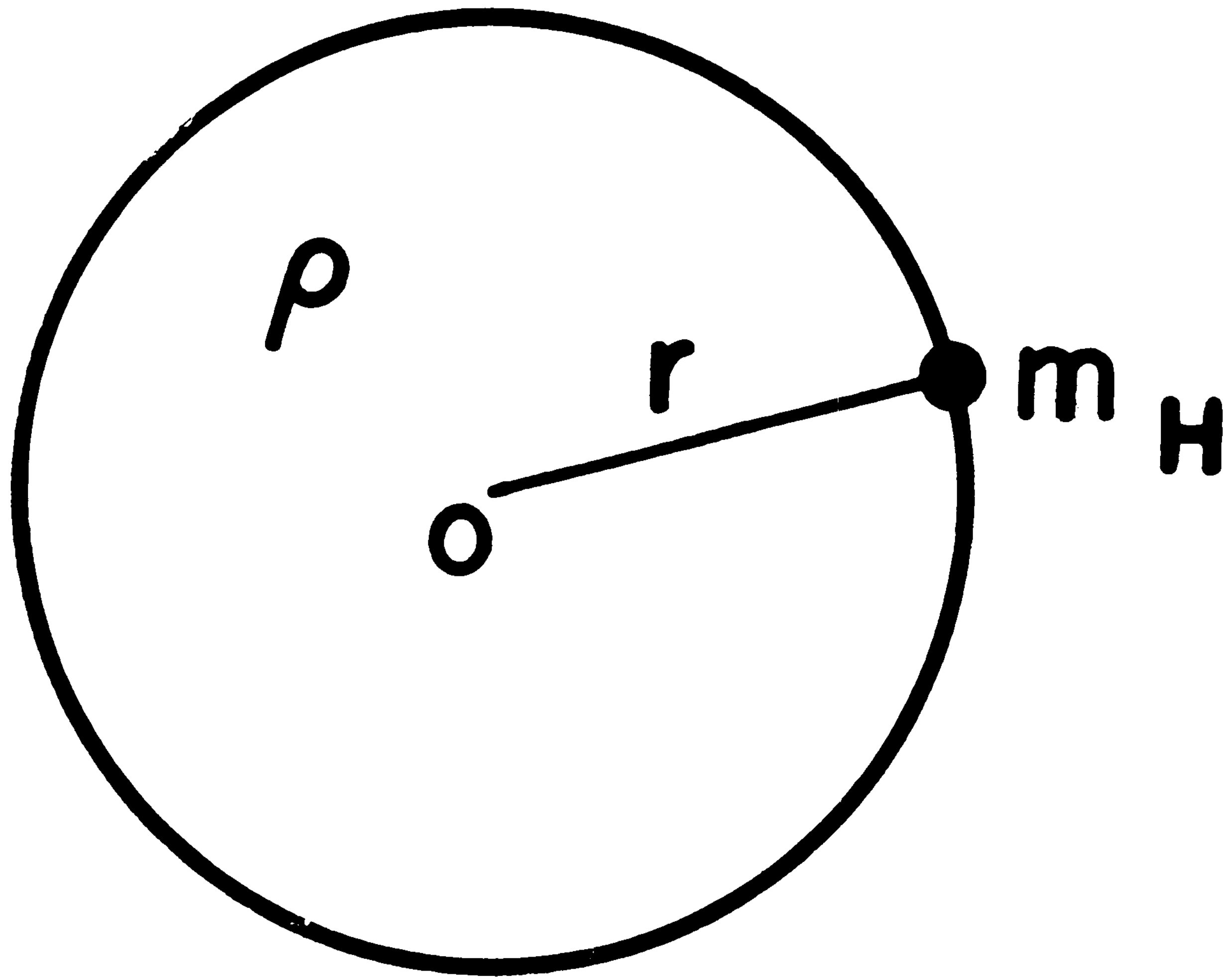


Figure 1. Model of a uniformly charged universe.

$$F = F_{el.} - F_{grav.} = (\mu - 1)\, F_{grav.}$$

$$F = (\mu - 1)\, \frac{4}{3}\, \pi\rho\, m_H\, Gr = kr \tag{15}$$

If $\mu - 1 > 0$, there will be a repulsive force which is proportional to r and which will lead to an expansion of the universe.

In order to achieve constant matter density in the universe despite the expansion, Lyttleton and Bondi propose the continuous creation of matter (hydrogen atoms) and hence necessarily also then the creation of charge. They propose a modification of Maxwell's equations to allow for the nonconservation of charge and solve the problem of a steady-state expanding universe with mass and charge creation. They obtain the following relationship between the mass density ρ, the Hubble constant T^{-1}, and the rate of matter creation Q:

$$\rho = \frac{1}{3}\, m_H\, Q\, T. \tag{16}$$

Using $T = 3 \times 10^{17}$ sec and $\rho = 10^{-29}$ gm/cm^3, they obtain

$$Q = 6 \times 10^{-23}\ \frac{\text{H atoms}}{\text{cm}^3\text{-sec}}$$

which corresponds to a creation rate of one hydrogen atom per second in a cube of 250 kilometers on an edge.

With constant matter density ρ the repulsive force given in equation (15) is consistent with a velocity which increases linearly with distance

$$v = \sqrt{K}\ r \tag{17}$$

where

$$K = (\mu - 1)\,\frac{4}{3}\,\pi\rho\,G$$

The observed expansion of the universe is

$$v = r/T \tag{18}$$

Equating (17) and (18) gives

$$T = \frac{1}{\left[(\mu - 1)\frac{4}{3}\pi\rho\,G\right]^{\frac{1}{2}}} \tag{19}$$

and hence $\mu = 5$

and $y = 2 \times 10^{-18}$. (20)

This is the charge inequality that Lyttleton and Bondi proposed to explain the observed expansion of the universe with a theory in which they allow for charge creation and a modification of Maxwell's equations. They also formulated their theory in the more general terms of de Sitter space-time to satisfy the cosmological principle that the universe appears the same as viewed from any position. The more general theory introduced no essential modifications of the basic conclusions of the Newtonian picture.

When ionization occurs, electrically neutral units will grow from the background of smoothed-out, un-ionized matter. These units are identified with galaxies or clusters of galaxies. Ions--primarily protons--which are expelled from these units by the electrostatic forces are identified with the hard component of the cosmic rays.

Hoyle [(7)] pointed out an error in the treatment of the modified Maxwell theory of Lyttleton and Bondi. The principal difference in conclusion reached by Hoyle is that the potential due to a charge will be of the form

$$\varphi = \frac{e}{r} \cos \left[(-\lambda)^{\frac{1}{2}} r\right] \tag{21}$$

where r is the distance from the charge and λ is a cosmological quantity

$$(-\lambda)^{\frac{1}{2}} \approx \frac{1}{\text{Radius of the Universe}}$$

(7) F. Hoyle, Proc. Roy. Soc. A 257, 431 (1960).

From equation (21) it is clear that the potential will change sign at sufficiently large distances, and thus the force between two like charges will change from repulsive to attractive.

Hoyle's interpreation then is that the electrostatic force would not be repulsive on a cosmological scale and lead to an expansion of the universe in the manner Lyttleton and Bondi proposed, but would rather be primarily attractive. Hoyle noted however that if matter and antimatter are both created at the same rate, if a hydrogen atom has a charge ye, and an antihydrogen atom a charge $-ye$, and if matter and antimatter become sufficiently separated, then repulsion of matter and antimatter will occur according to equation (21) and expansion of the universe would occur. Hoyle's theory also requires that $y \simeq 2 \times 10^{-18}$.

3. Experimental Evidence on Charge Difference.

Now I would like to discuss what terrestrial laboratory experiments have established about the electron-proton charge difference.

One of the earliest experiments was the Millikan oil drop experiment [8]. Millikan studied the motion of droplets of various liquids which had been charged by different means

(8) R. A. Millikan, The Electron (University of Chicago Press, Chicago, 1917), 1st ed. pp. 80-83.

such as by friction, by use of x-rays, or by capture of ions from the air. From the observation of the motion of these droplets under the forces of gravity, of viscous drag, and of an electric field, Millikan was able to show that in all cases every droplet had a charge which was an integral multiple of the smallest unit. He studied charges of both signs and he found that

$$\frac{\text{positive charge unit}}{\text{negative charge unit}} = 1 \pm 1/1500$$

A macroscopic interpretation of this result can be given in terms of the electron-proton charge difference [9]. A typical oil droplet is a sphere with a radius of about 10^{-4} cm and a density of 1 gm/cm^3. The number, N, of proton-electron pairs in one of these droplets is then $N \simeq 2.5 \times 10^{12}$. Millikan's observations require that

$$Nye < e/1500$$

and hence

$$y < 3 \times 10^{-16}$$

Another macroscopic experiment by a gas efflux method

(9) V. W. Hughes, Phys. Rev. 76, 474 (1949) (A), ibid., 105, 170 (1957).

was first done by Piccard and Kessler [4] and will be discussed later.

I should like to discuss next an atomic beam experiment which has recently been done by Zorn, Chamberlain, and Hughes [10, 11, 9]. The method of the experiment is to study the deflection of a molecular beam in a homogeneous electric field. If an atom is neutral, it will not be deflected, but if there were a difference between the electron and proton charge magnitudes then an atom would have a net charge and it would be deflected.

We used a classic molecular beam technique [12] as illustrated in Figure 2.

---

(10) J. C. Zorn, G. E. Chamberlain and V. W. Hughes, Bull. Am. Phys. Soc. 6, 63 (1961); Proceedings of the Tenth Annual International Conference on High Energy Physics (Interscience Publishers, New York, 1960), p. 790.

(11) J. C. Zorn, G. E. Chamberlain and V. W. Hughes, Bull. Am. Phys. Soc. 5, 36 (1960).

(12) P. Kusch and V. W. Hughes, "Atomic and Molecular Beam Spectroscopy" in Handbuch der Physik 37/1. S. Flügge, ed. (Springer-Verlag, Heidelberg, 1959), p. 6.

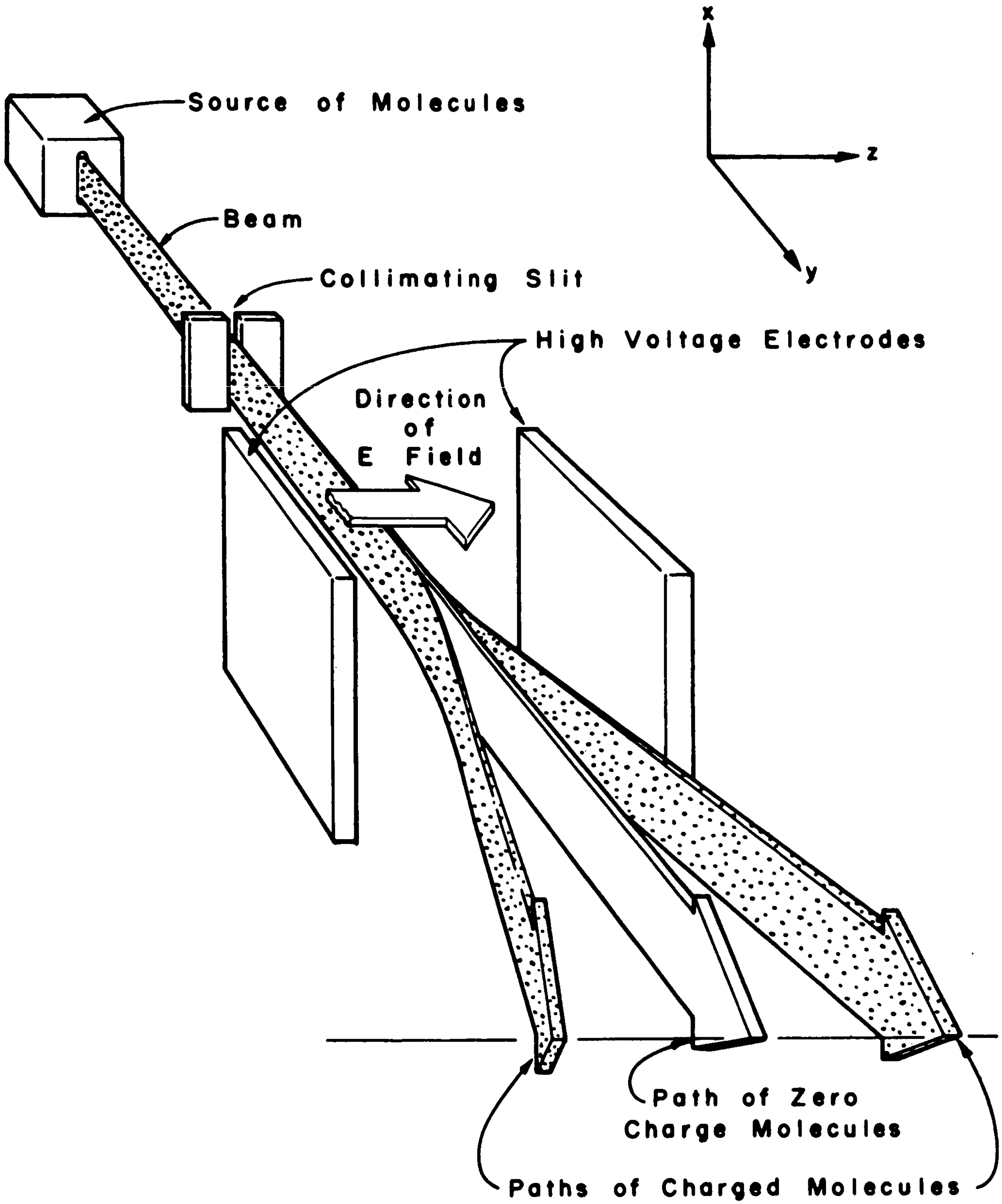


Figure 2. Molecular beam measurement of atomic or molecular charge.

The beam is defined by a source slit and a collimating slit so that it has a ribbon-like cross section which is narrow in the transverse horizontal direction and long in the vertical direction. This beam passes through a homogeneous electric field which would deflect the beam if the atoms were charged.

Figure 3 shows a horizontal cross section of the apparatus in greater detail. In terms of the geometry of Figure 3, the deflection that a charged molecule of velocity v would experience due to the electric field is given by

$$s_v = \frac{qE}{2mv^2} L_1 (L_1 + 2L_3) \tag{22}$$

where q, m, and v are the charge, mass and velocity of the particle in the beam and E is the electric field strength. In particular, a molecule with the most probable velocity α of molecules in the source $(\alpha = \sqrt{2kT/m})$ is deflected by the amount

$$s_\alpha = \frac{qE}{4kT} L_1 (L_1 + 2L_2) \tag{23}$$

where T is the source temperature and k is Boltzmann's constant.

In our recent experiment

$$L_1 = 200 \text{ cm}, \quad L_2 = 30 \text{ cm}$$

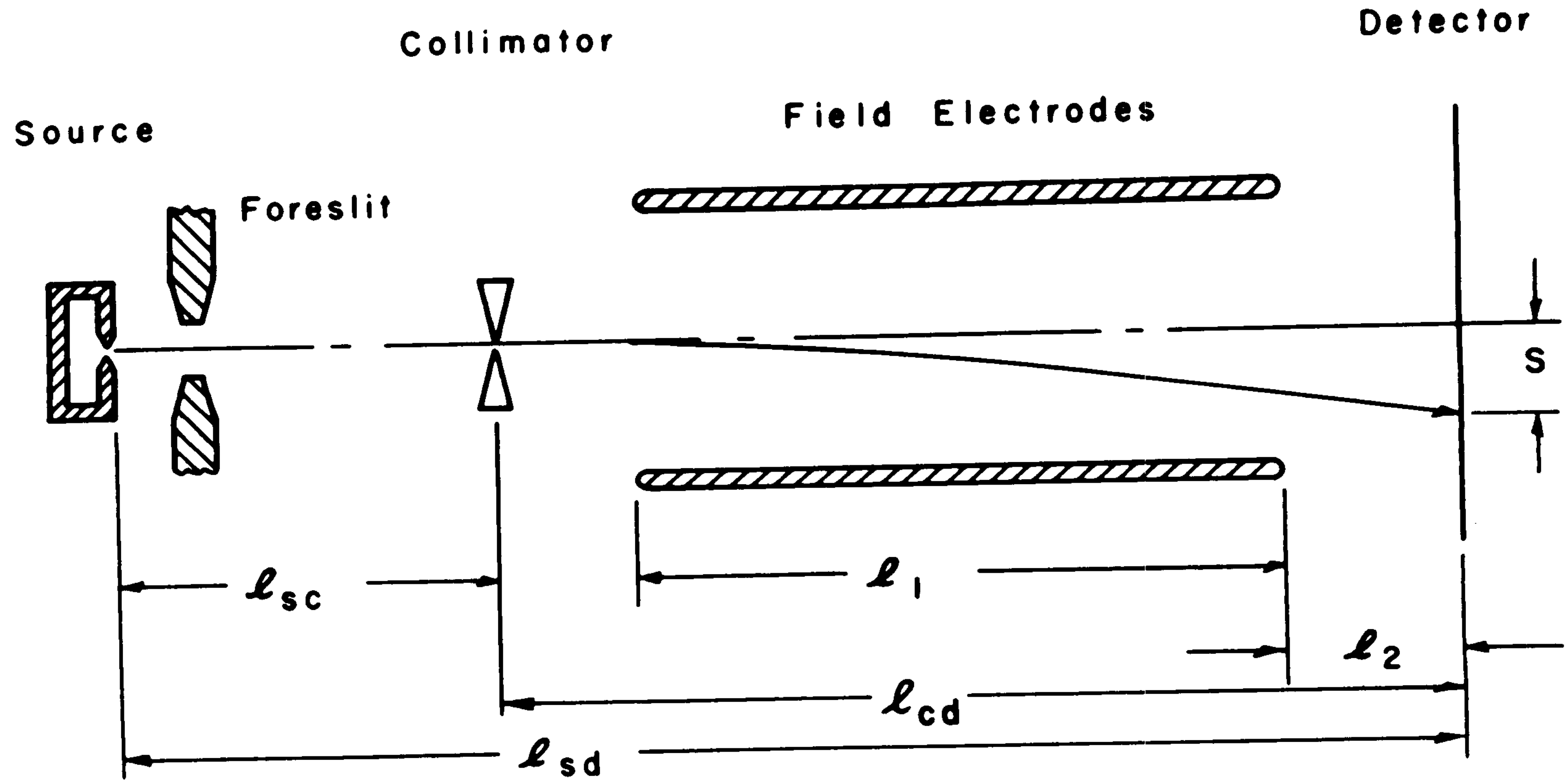


Figure 3. Geometry of apparatus, showing trajectory of an atom which has been deflected by the electric field.

and $$E = 10^5 \text{ volts/cm.}$$

The experiment was done for cesium and potassium atoms and the oven temperature was about 500 $^{\circ}$K. Our detector sensitivity was such that a deflection of 10^{-5} cm could be detected. Hence the minimum detectable atomic charge was

$$q \simeq 3 \times 10^{-17} e$$

For cesium the atomic number is 55; so the minimum detectable charge on an electron-proton pair, δq , is smaller by a factor of 55:

$$\delta q \sim 6 \times 10^{-19} e$$

This sensitivity is in the range of interest for the Lyttleton-Bondi theory.

There are some complications which are important to the experiment. Because of the smallness of the deflections being observed, electric field inhomogeneities can produce comparable deflections associated with the polarization of the atoms. The atoms have no permanent electric dipole moments, but in an electric field an electric dipole moment is induced. If the field is inhomogeneous, there will be a force on this induced electric dipole moment. In our experiment such field inhomogeneities arise at the ends of the field region. If the energy of the atom in the field is $W(E)$, then the force due to the induced dipole moment is

$$\vec{F} = -\nabla W(E) = -\frac{\partial W}{\partial E} \nabla |E| \qquad (24)$$

It is apparent from the form of equation (24) that the direction of the force does not change with the direction of the field. Hence by reversing the polarity of the potential across the electrodes, we can distinguish between this dipole polarizability force and the force on a net atomic charge.

Another complication in interpreting the deflection measurements is the spread in velocities of the atoms. The velocity distribution is a Maxwellian one for particles effusing through an opening in the oven:

$$I_v \, dv = \frac{2I}{\alpha^4} v^3 e^{-v^2/\alpha^2} \, dv \qquad (25)$$

where I is the total beam intensity. The observed deflection is given by an average over this velocity distribution.

Figure 4 illustrates a third complicating factor which must be considered. The source and detector slits have finite widths, so that we obtain a beam intensity distribution in the detector plane which is trapezoidal. In addition, the detector has a finite width.

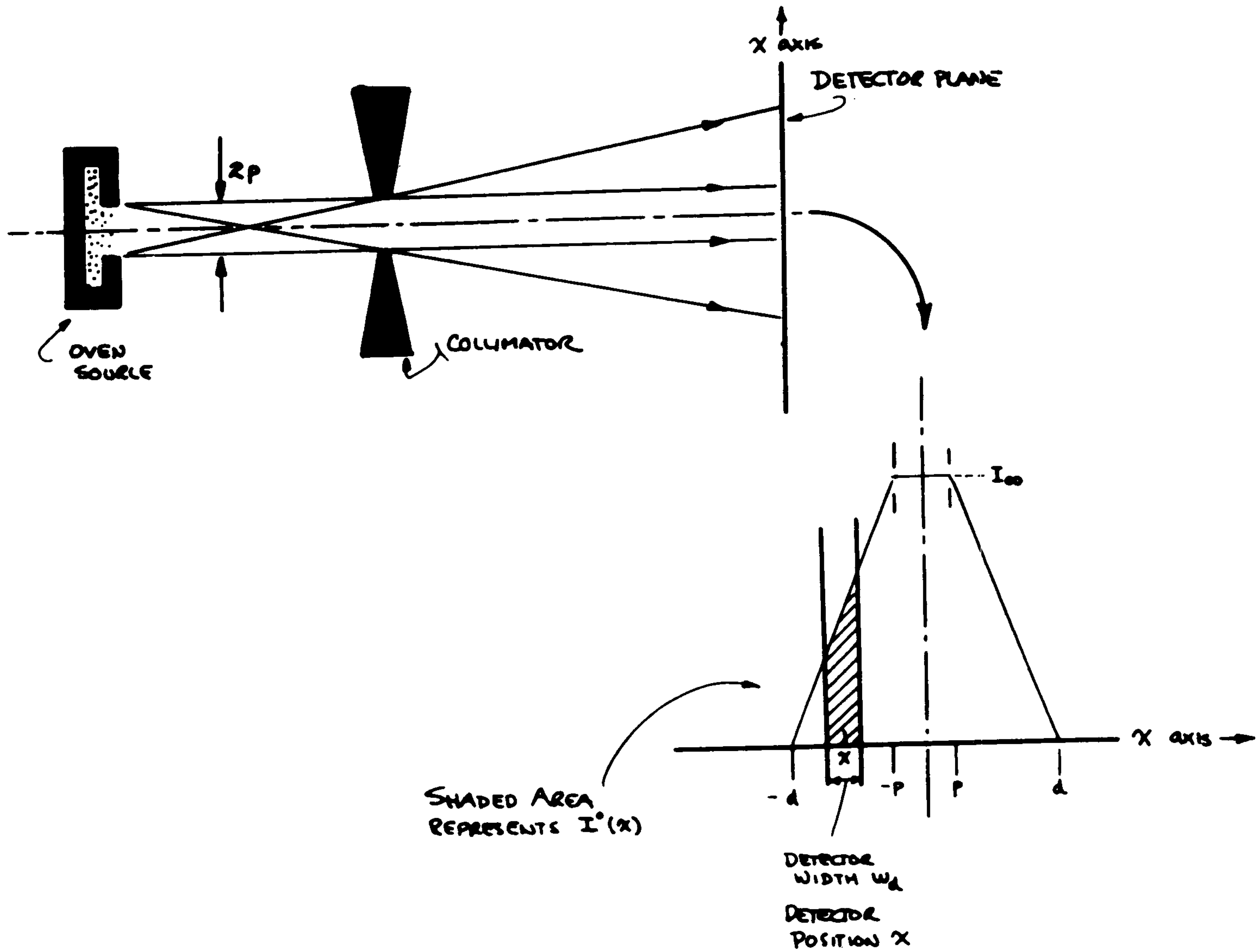


Figure 4. Effect of finite slit widths on beam intensity distribution.

In order to relate the observed intensity pattern to s_α, it is necessary to integrate over the width of the beam path and over the velocity distribution. The relation between s_α and the change in intensity with the detector positioned where the beam intensity has one half its maximum value is given by:

$$\frac{\Delta I}{I} = \frac{2s_\alpha}{d - p} \tag{26}$$

where d is the half width of the penumbra of the beam in the detector plane and p is the half width of the umbra. The analysis has also been done in another way which does not require an a priori knowledge of the slit geometry and alignment but uses only the observed beam intensity distribution.

Some technical features of the experiment and of the apparatus will now be discussed. The choice of the atom is dictated largely by atomic beam technology. The only property of the atom that appears in the deflection equation (23) is the temperature at which it must be produced. This should be as low as possible. For this experiment we desire an atom containing many electron-proton pairs. Alkali atoms are used because they are produced conveniently in beams at relatively low temperatures and they are detected efficiently with a hot wire surface ionization detector. Figure 5 shows the oven used to produce the beam of potassium or cesium atoms. It is used at a temperature of about 500° K.

Figure 6 shows the observed and calculated beam intensity distribution with oven and collimator slit widths of 0.004 cm. The detector width is also 0.004 cm. The agreement between the two curves is good; the small discrepancy is attributed to atomic beam scattering, slit misalignment, and imperfect knowledge of slit dimensions. The detector is placed at one of the two half-maximum intensity points in order to obtain the maximum change in intensity for a given s_{α} .

Figure 7 shows the electric field assembly in vertical cross section. The parallel plates are made of aluminum and are about two meters in length with a spacing of 1 or 2 mm. Electric fields of 100 kv/cm are obtained before breakdown occurs.

Figure 8 shows some of the observed data. The change in beam intensity Δ observed with the detector placed at the two half-maximum intensity points (z_1 and z_2) is plotted as a function of electric field for both polarities of the field (A and B).

The deflection of the beam due to a net atomic charge

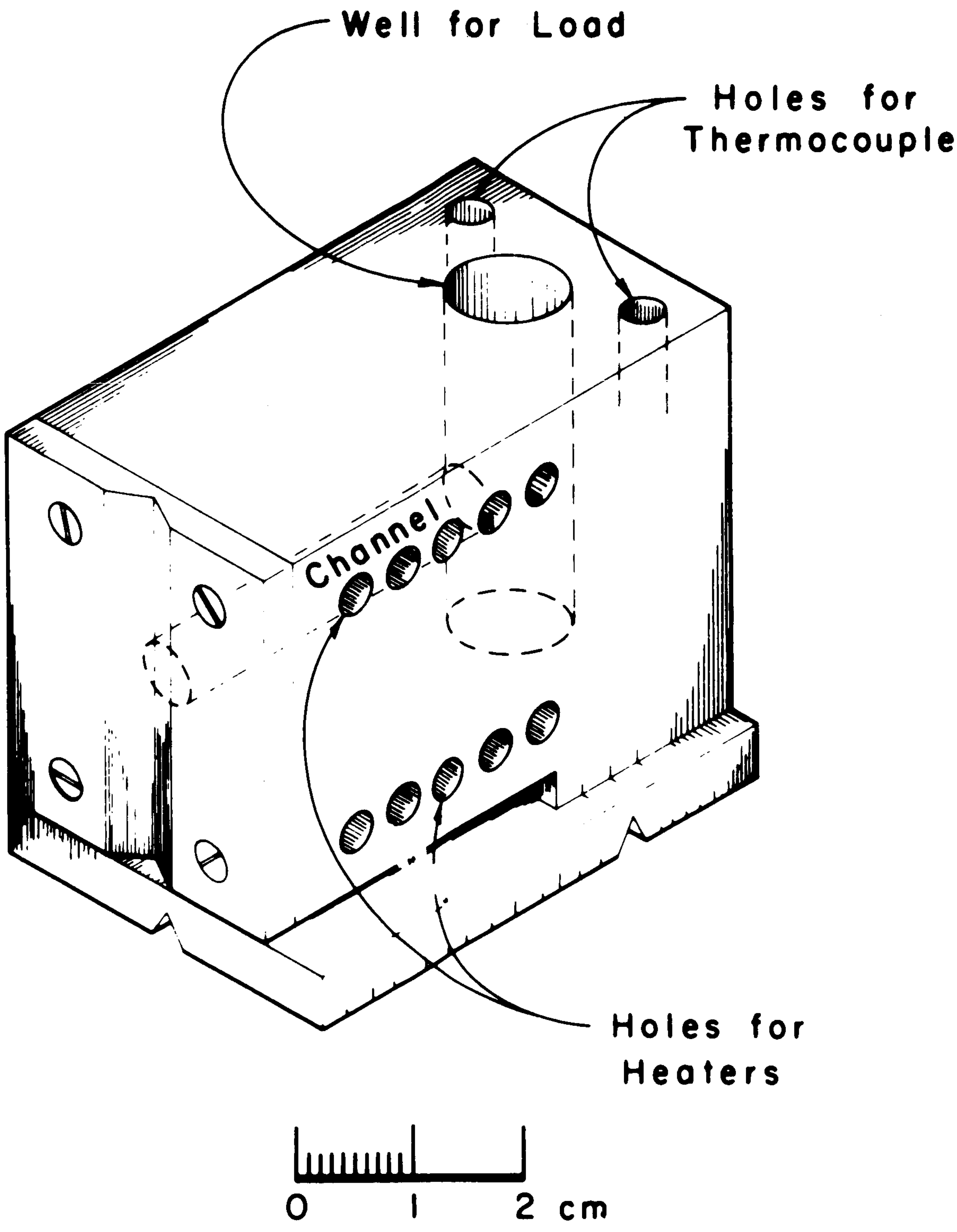


Figure 5. Conventional oven used for alkali atoms.

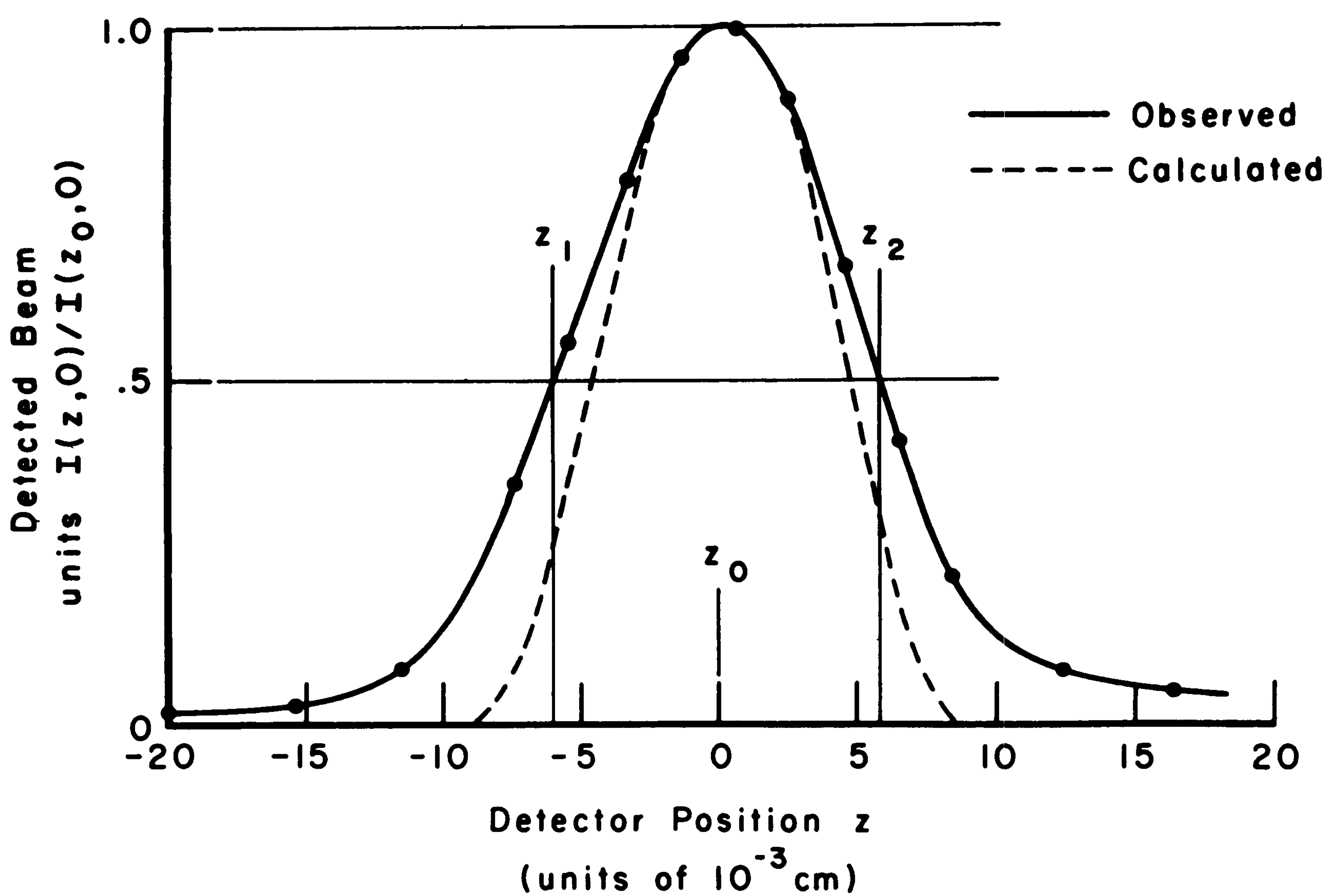


Figure 6. Theoretical beam shape for classical trajectories and ideal geometry compared to an observed zero-field intensity distribution for potassium atoms.

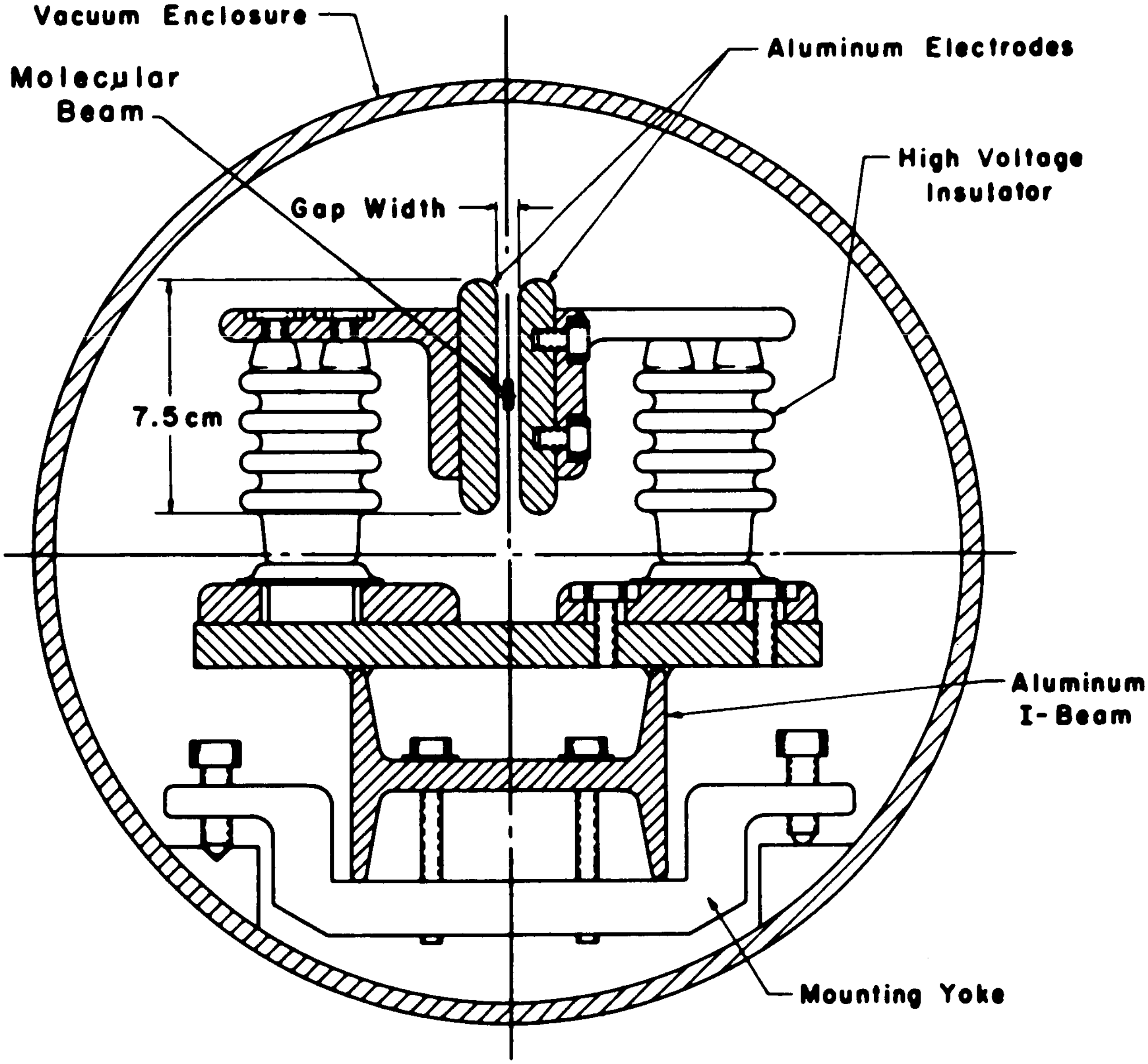


Figure 7. Cross-section of electrode assembly.

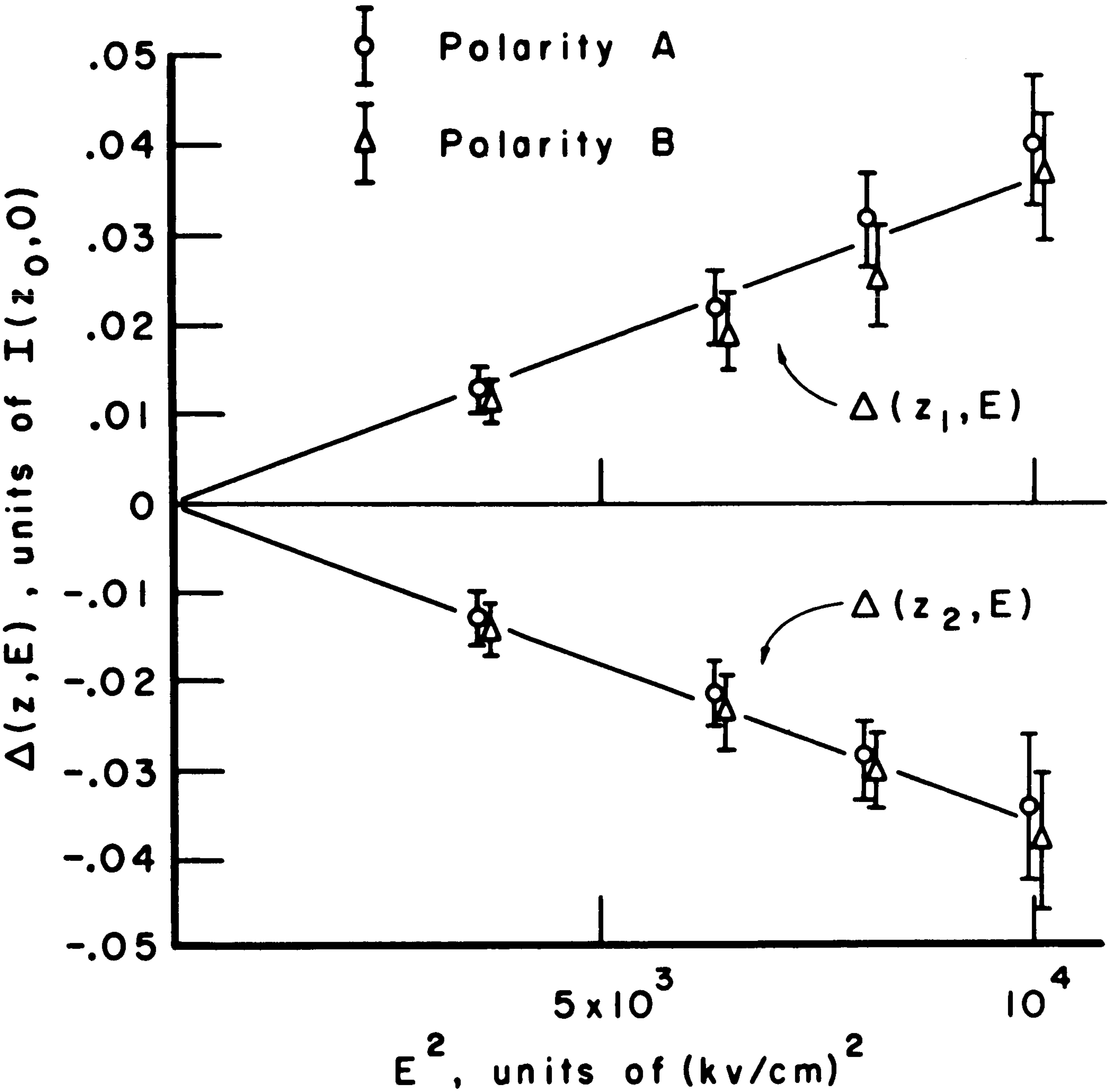


Figure 8. Beam intensity as a function of electric field at half-maximum intensity points (z_1 and z_2).

is directly proportional to E and, at the field strengths used in this experiment, the deflection from the induced dipole moment is proportional to E^2. The observed dependence of $\Delta(z_i,E)$ is shown in Figure 8. It is seen that $\Delta(z_i,E)$ is linearly proportional to E^2 up to a field E of about 10^5 v/cm, as expected for deflections due to dipole polarizability alone. At still higher fields Δ is no longer proportional to E^2; indeed both $\Delta(z_1,E)$ and $\Delta(z_2,E)$ decrease with an increase of E at sufficiently high values of E. This behavior is not consistent with deflection due to a net atomic charge and a dipole polarizability but rather is explained by an attenuation of the atomic beam at the higher fields. The beam appears to be attenuated in proportion to the gap current, and this gives rise to a field dependent signal change $D(z_i,E)$ not associated with an electric deflection of the beam atoms.

Table I shows the results deduced from such measurements on Potassium and Cesium atoms and on hydrogen and deuterium molecules. The upper limits for the charges are given. The upper limits on the charge are considerably higher for hydrogen and deuterium than for the alkalis. This is due to the fact that the Pirani detector for hydrogen is not as efficient as

TABLE I.

RESULTS	INTERPRETATION
Gases:	
$q(H_2) < \|2\| \times 10^{-15}\ q_e$	$\delta q < \|1\| \times 10^{-15}\ q_e$
$q(D_2) < \|2.8\| \times 10^{-15}\ q_e$	$q_n < \|1.7\| \times 10^{-15}\ q_e$
Alkalis:	
$q(K) = (-3.8 \pm 11.8) \times 10^{-17}\ q_e$	$q(K) = 19\ \delta q + 20\ q_n$
$q(Cs) = (+1.3 \pm 5.6) \times 10^{-17}\ q_e$	$q(Cs) = 55\ \delta q + 78\ q_n$
	If the value of δq is considered to be independent of q_n, the above are a pair of equations in two unknowns; solution gives the limits
	$\delta q = (-.85 \pm 2.7) \times 10^{-17}\ q_e$
δq: electron-proton charge difference q_n: neutron charge q_e: absolute value of electron charge	$q_n = (+.61 \pm 2.0) \times 10^{-17}\ q_e$
	But the neutron decay: $n \rightarrow p + e^- + \bar{\nu}$ indicates $\delta q = q_n$, so
	$\delta q = (1.0 \pm 4.2) \times 10^{-19}\ q_e$

the hot wire surface ionization detector for the alkalis so that the gas apparatus was shorter and less sensitive to small deflections than the alkali apparatus.

The charge of an atom or molecule is assumed to be completely given by the scalar sum $q = Z\delta q + Nq_n$, where Z is the number of electron-proton pairs, $\delta q = q_p - q_e$ is the electron-proton charge difference, N is the number of neutrons, and q_n is the neutron charge. The most direct determination of a limit for δq is obtained from the measurement of the net charge of the hydrogen molecule:

$$\left|\delta q\right| = \frac{\left|q\ (H_2)\right|}{2} < 1 \times 10^{-15}\ q_e \qquad (27)$$

In addition, the result from deuterium gives a limit for q_n :

$$q_n < 2.4 \times 10^{-15}\ q_e \qquad (28)$$

Smaller limits than the above can be obtained from the experimental values for the charges of cesium and potassium.

$$q(Cs) = 55\ \delta q + 78\ q_n = (13 \pm 56) \times 10^{-18}\ q_e \qquad (29)$$

$$q(K) = 19\ \delta q + 20\ q_n = (-38 \pm 118) \times 10^{-18}\ q_e \qquad (30)$$

As simultaneous equations in δq and q_n , the solution gives

$$\delta q = (-8.5 \pm 27) \times 10^{-18}\, q_e \tag{31}$$

independently of the value of q_n , and

$$q_n = (6.1 \pm 20) \times 10^{-18}\, q_e \tag{32}$$

independently of the value of δq.

A still smaller limit for the electron-proton charge difference can be given if one assumes that $\delta q = q_n$. This relation follows from the usual assumption that charge is conserved in beta decay of the neutron $(N \rightarrow p + e + \bar{\nu})$ and that the charge of the antineutrino is zero (*). Then $\delta q = q\ (\text{atom})/(Z + N)$ and we obtain from $q(\text{Cs})$:

$$\delta q = (1.0 \pm 4.2) \times 10^{-19}\, q_e \tag{33}$$

With improved vacuum, electric field conditions, and detector stability we believe our atomic beam experiment on the alkalis could be improved in sensitivity by about a

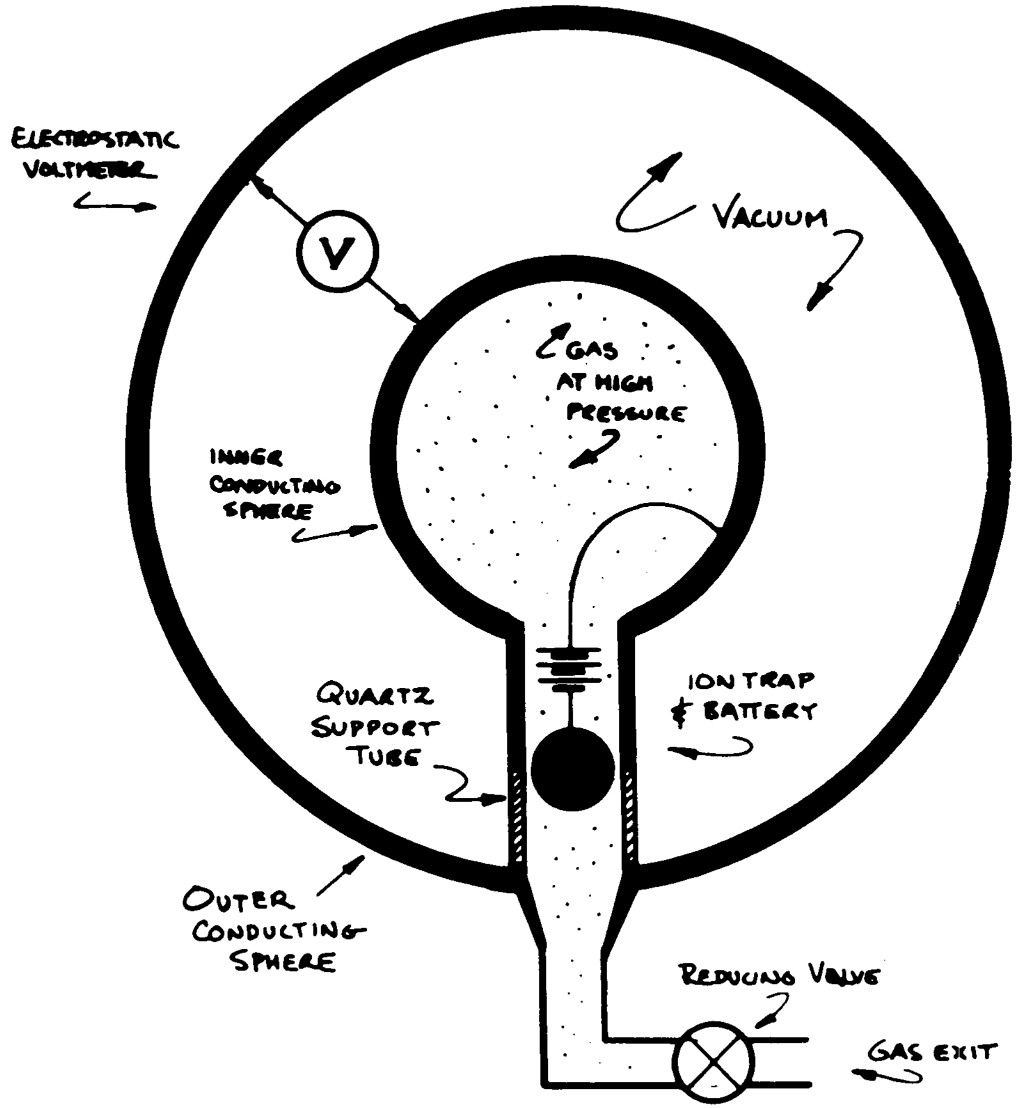
---

(*) An upper limit to the neutrino charge can be obtained by considering that the neutrino is a Dirac particle with a mass of 500 ev (upper limit to the allowed neutrino mass) and computing the upper limit to the charge that is consistent with neutrino cross-section data (J. S. Allen, The Neutrino (Princeton University Press, Princeton, 1958). The limit found for the neutrino charge in this way is about $10^{-10}\, q_e$.

factor of 100. An atomic beam experiment on thermal neutrons was done by Shapiro and Estulin who obtained an upper limit for the neutron charge of 6×10^{-12} q_e .

I would like to discuss briefly the macroscopic gas efflux experiment done first by Piccard and Kessler [4], which measures the total charge Q of M gas molecules by observing the change in potential of a metal container relative to its surroundings when gas effuses from the container. Figure 9 shows their apparatus consisting of two concentric conducting spheres which form a spherical capacitor. The inner sphere can be filled with a gas. The voltage between the two spheres depends on the capacity, on the surface charge on the inner sphere, and on the volume charge carried by the gas.

Piccard and Kessler filled the inner sphere with 20 to 30 atmospheres of CO_2 or N_2. Then they allowed the gas to effuse from the inner sphere and measured the change in potential across the capacitor. If the gas were neutral and there were no changes in the dimensions of the sphere, then there should be no change in the potential. On the other hand, if the gas had a net charge due to a proton-electron charge difference, then the potential would change when the



$$\Delta Q = C \Delta V$$

CO_2: PICCARD & KESSLER, ARCH SCI PHYS ET NAT 7, 340 (1925)
N_2, A HILLAS & CRANSHAW, NATURE 184, 892 (1959).

Figure 9. Piccard-Kessler apparatus for gas efflux experiment.

gas leaves the inner sphere. The efflux of ions or electrons was prevented, or at least made difficult, by biasing a small obstacle in the throat of the exhaust tube relative to the inner sphere such that ions are trapped in the inner sphere and are not exhausted with the neutral gas. From their measurements they determined that $\delta q \leq 5 \times 10^{-21}$ e.

Figure 10 shows a modern version of this same experiment by King [13,14]. King did his experiment with hydrogen and on helium.

Conservatively we can interpret his results as setting an upper limit for the charge on H_2 of less than $10^{-19}\ q_e$.

A modern extension of Millikan's oil drop experiment using a small, magnetically suspended metal sphere has been proposed to achieve a higher sensitivity in the determination of δq

Table II presents a summary of experimental information on the electron-proton charge difference.

4. Interpretation of Results.

The atomic beam deflection experiment on the alkali atoms

---

(13) J. G. King, Phys. Rev. Let. 5, 562 (1960).

(14) A. M. Hillas and T. E. Cranshaw, Nature 184, 892 (1959), ibid.186, 459 (1960). H. Bondi and R. A. Lyttleton, Nature 184, 974 (1959).

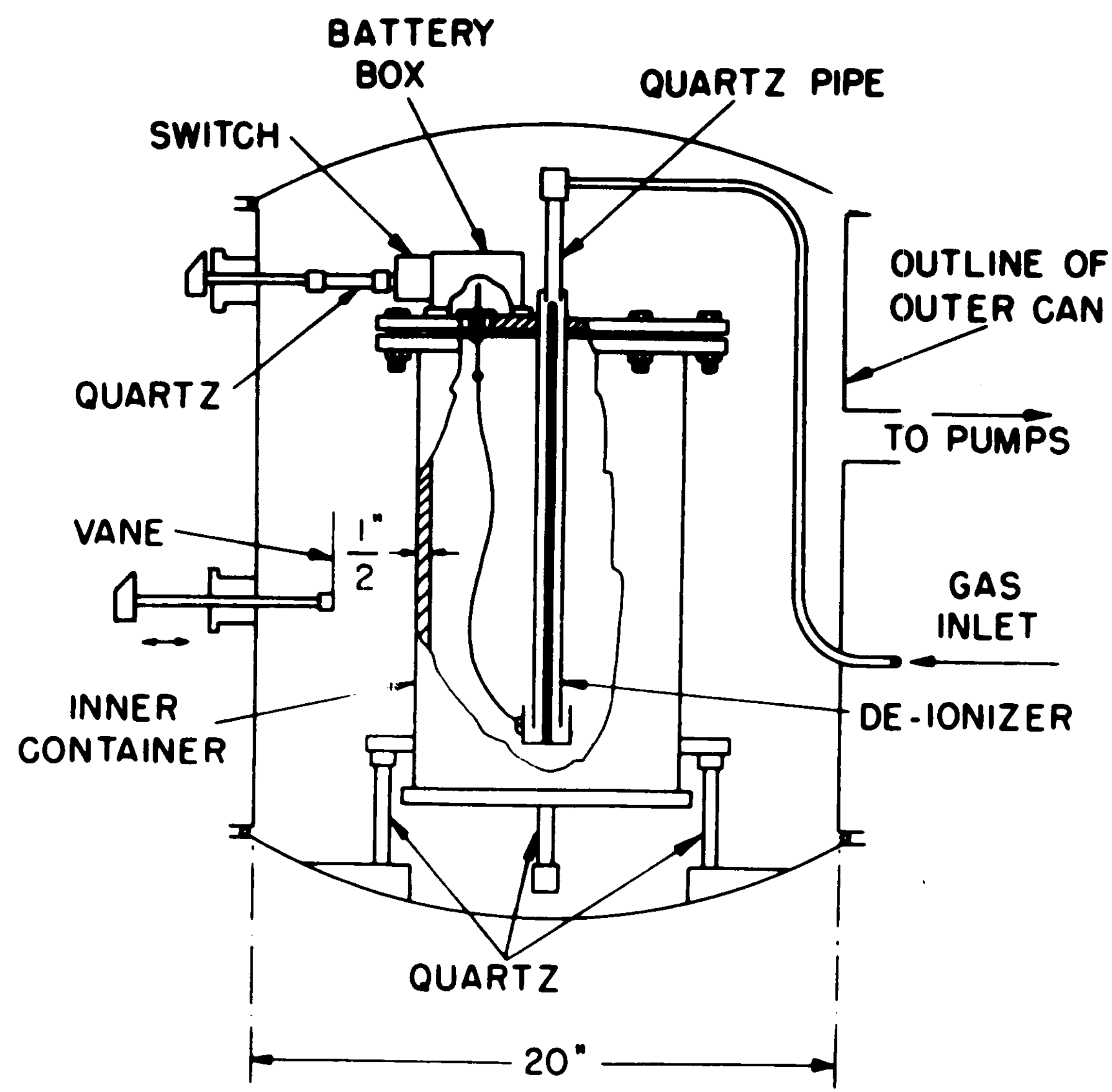


Figure 10. King apparatus for gas efflux experiment.

TABLE II.

Measured Limits for the Charge of Molecules

Method	Molecule	q(Molecule) in units of q_e	$\delta q = q(\text{molecule})/(Z + N)$	Investigators
	CO_2	$<2.2 \times 10^{-19}$	$<5 \times 10^{-21}$	Piccard & Kessler[a]
By	A	$(4 \pm 4) \times 10^{-20}$	$(1 \pm 1) \times 10^{-21}$	Hillas & Cranshaw[b]
Gas	N_2	$(6 \pm 6) \times 10^{-20}$	$(2.1 \pm 2.1) \times 10^{-21}$	Hillas & Cranshaw
Efflux	H_2	$(-2.5 \pm 1.5) \times 10^{-20}$	$(-1.3 \pm 0.8) \times 10^{-20}$	King[c]
	He	$(4 \pm 2) \times 10^{-20}$	$(1 \pm 0.5) \times 10^{-20}$	King
	CsI	$<4 \times 10^{-13}$	$<1.5 \times 10^{-15}$	Hughes[d]
	Free neutron	$<6 \times 10^{-12}$		Shapiro & Estulin[e]
By	CsF	$<2 \times 10^{-14}$	$<1.3 \times 10^{-16}$	Zorn, Chamberlain & Hughes[f]
Beam	KF	$<1 \times 10^{-13}$	$<1.7 \times 10^{-15}$	Zorn, Chamberlain & Hughes[f]
Deflection	H_2	$<2 \times 10^{-15}$	$<1 \times 10^{-15}$	Present work
	D_2	$<2.8 \times 10^{-15}$	$<0.7 \times 10^{-15}$	Present work
	K	$(3.8 \pm 11.8) \times 10^{-17}$	$(1 \pm 3) \times 10^{-18}$	Present work
	Cs	$(1.3 \pm 5.6) \times 10^{-17}$	$(1 \pm 4.2) \times 10^{-19}$	Present work

By Mossbauer effect, a limit of 1×10^{-15} q_e for the charge of the photon has been established[g]

[a] Reference 6 [c] Reference 8 [e] Reference 10

[b] Reference 7 [d] Reference 9 [f] Reference 11

[g] Grodzins, Engelberg & Bertozzi, Bull. Am. Phys. Soc. 6, 63 (1961)

provides a limit for δq of $5 \times 10^{-19} q_e$. This limit is about 1/4 the value of δq required by the theory of the expanding universe proposed by Lyttleton and Bondi. Furthermore, the macroscopic experiments by the gas efflux method provide the even smaller limit of $10^{-21} q_e$ to $10^{-20} q_e$. All of these results provide strong evidence against the form of the Lyttleton-Bondi proposal which requires $\delta q = 2 \times 10^{-18} q_e$; they do not test the alternative, though less attractive, form of the Lyttleton-Bondi proposal which requires a greater number of protons than electrons in the universe.

The equality of the electron and proton charge magnitudes has been established with unusually high precision in this and other recent experiments; hence they offer no support for the suggestion that baryon conservation might be simply a consequence of charge conservation. Furthermore, it would seem that any theory of elementary particles should require that the renormalized electron and proton charge magnitudes be equal.

REFERENCES

1. M. Gell-Mann, Proceedings of the Tenth Annual International Conference on High Energy Physics (Interscience Publishers, Inc., New York (1960), p. 792.

2. G. Feinberg and M. Goldhaber, Proc. Natl. Acad. Sci. U. S. 45, 1301 (1959).

3. W. F. G. Swann, Phil. Mag. 3, 1088 (1927); Astrophysical J. 133, 733 (1961).

4. A. Piccard and E. Kessler, Arch. Sci. Phys. et Nat. (Geneva) 7, 340 (1925).

5. P. M. S. Blackett, Nature 159, 658 (1947).

6. R. A. Lyttleton and H. Bondi, Proc. Roy. Soc. A 252, 313 (1959).

7. F. Hoyle, Proc. Roy. Soc. A 257, 431 (1960).

8. R. A. Millikan, The Electron (University of Chicago Press, Chicago, 1917), 1st ed. pp. 80-83.

9. V. W. Hughes, Phys. Rev. 76, 474 (1949)(A), ibid., 105, 170 (1957).

10. J. C. Zorn, G. E. Chamberlain and V. W. Hughes, Bull. Am. Phys. Soc. 6, 63 (1961); Proceedings of the Tenth Annual International Conference on High Energy Physics Interscience Publishers, New York, 1960), p.790.

11. J. C. Zorn, G. E. Chamberlain and V. W. Hughes, Bull. Am. Phys. Soc. 5, 36 (1960).

12. P. Kusch and V. W. Hughes, "Atomic and Molecular Beam Spectroscopy" in Handbuch der Physik 37/1. S. Flügge, ed. (Springer-Verlag, Heidelberg, 1959), p. 6.

13. J. G. King, Phys. Rev. Let. 5, 562 (1960).

14. A. M. Hillas and T. E. Cranshaw, Nature 184, 892 (1959). ibid., 186, 459 (1960). H. Bondi and R. A. Lyttleton, Nature 184, 974 (1959).

QUANTIZATION OF GENERAL RELATIVITY

Lecture XIV

J.L. Anderson

Stevens Institute of Technology

Seminar on Gravitation and Relativity, NASA Goddard Space Flight Center, Institute for Space Studies, New York, N.Y.; edited by H.Y. Chiu and W. F. Hoffmann.

Preface

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R. H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution of space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a certain amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and correct the final manuscript. Hence, we accept responsibility for errors and omissions.

H. Y. Chiu

W. F. Hoffmann

Since the early days of quantum mechanics many people have felt that a complete quantized theory of matter must include the theory of gravitation. For this reason, there has been much effort to join together these two fundamental aspects of the physical world. One of the earliest attempts at this union is the program outlined by Bergmann[1] in 1947. Since then many other approaches have been taken.[2]

Studies on the quantization of general relativity to date have concerned themselves mainly with a better understanding of the classical formulation of general relativity as a prelude to applying one of the several techniques of quantization to it. Historically, and most commonly, the quantization of a given classical system proceeds from the Hamiltonian formulation of that theory. Given the canonical coordinates and momenta which describe a state of the system and the Hamiltonian as a function of these variables, there is a more or less unique algorithm for constructing the corresponding quantum description of the system. Thus much of the effort in quantum relativity has been towards constructing a Hamiltonian formulation of general relativity.

However, because of difficulties inherent in a Hamiltonian formulation of general relativity, other procedures of quantization are being tried. One such procedure, under current investigation by Bergmann and Komar[3] and also DeWitt[4], is to look at the coordinate invariant quantities one can construct in the theory and calculate their commutators in the classical theory. Once these commutators are known for a complete set of invariants, one can hope to find an operator representation for them which reproduces their classical commutator algebra. DeWitt has employed a generalization of the method of constructing commutators developed by Peierls[5] while Bergmann and Komar have made extensive use of the theory of infinitesimal canonical transformations.

Another approach, currently being worked on by Wheeler and his group[6], is that of the Feynman path integral formulation of quantum mechanics. The Schwinger variational principle has also been applied to general relativity by Arnowitt and Deser.[7]

These alternate approaches to the usual Hamiltonian quantization have all been initiated with the hope of overcoming the difficulties associated with Hamiltonian quantization. This goal has not been reached. The difficulties in Hamiltonian quantization reappear in one or another form. Since an exhaustive treatment of each of these methods of quantization as applied to general relativity would be impossible here, we will concentrate our attention mainly on the Hamiltonian form of quantization. This form contains all of the essential difficulties to be encountered in quantizing general relativity. Furthermore, this quantization scheme is the one we understand the best

II. Motives for Quantizing the Gravitational Field

Before I discuss the details of the difficulties, I would like to point out some of the pros and cons of such an undertaking.

It has been argued by many people that since the gravitational field is an extraordinarily weak field around 10^{-40} of the strength of the electromagnetic field, one should not expect to see any effects of gravitation on a microscopic atomic or nuclear level. Consequently, gravitation will play no essential role in elementary particle processes or any of the other microscopic phenomena we know about. For this reason many people do not believe in the necessity of quantizing gravitation field. However, there are a number of arguments that suggest that this argument based solely on the weakness of the gravitational field may be misleading

First of all the general theory of relativity is a non-linear theory and is intrinsically non-linear, unlike electrodynamics which only becomes non-linear through its coupling with the Dirac field. The gravitational field is non-linear even without coupling to some other source field. Thus there is no assurance that the concepts and procedures developed in electrodynamics are meaningful in the case of general relativity. We do not even know if the gravitational analogue of the photon exists.

Recently, however, Feynman has taken the position that it would be interesting to see how far one could get, by applying the concepts and procedures usually used in quantum field theories to general relativity, and by treating it as a linear theory with the non-linear part acting as an effective self-interaction. In this way, he has obtained the classical results of general relativity concerning the ~~true~~ three experimental tests of the theory. His result will be discussed in more detail in the last chapter of this lecture series.

On the other hand, one can argue that the full non-linearity is an essential feature of the problem and cannot be treated as a small perturbation It is possible that when one gets very close to an elementary particle the gravitational field becomes large enough that the non-linearities begin to play an essential role and begin to change the character of the problem in a qualitative manner. This corresponds to a situation in the theory of differential equations: In a non-linear system there exist solutions which cannot be reached by linear approximations.

There is an example of such a situation in classical field theory. This is the Born-Infeld theory of electrodynamics. Born and Infeld found an exact solution corresponding to a point charge In the full non-linear theory the self-energy of the charge is infinite and there is an automatic cut-off to interactions of the charge with electric fields of arbitrarily high frequency. The linear approximation to this theory is Maxwell electrodynamics where those results do not hold, even if non-linear terms are included as perturbations. This example shows that in some aspects of the theory one cannot expect qualitative similarities between a non-linear theory and its linearized version. This is directly related to the problem discussed by Wheeler in Lecture X. There he introduced non-Euclidean topology into the theory. As long as the topology is Euclidean, we are justified to make a linear approximation of the gravitational field equations with the non-linear term taken to be a small perturbation.

However, if one takes seriously the idea that in the neighborhood of elementary particles the topology may be different from Euclidean, then it is not possible to treat the gravitational field as a weak field. There is no suitable first-order approximation to the field. It is necessary to quantize the whole theory right at the beginning.

There are also several other arguments in favor of quantizing the gravitational field. It is believed that all particles produce gravitational fields. If these gravitational fields are effectively classical, then by measuring all components simultaneously can determine both the position and velocity of a particle simultaneously and thus violate the uncertainty principle. Hence the gravitational field must not be classical but must fluctuate in order to be compatible with quantum concepts.

Pauli argued that such fluctuations in the gravitational field may smear out the light cone. This in turn might conceivably furnish a natural cut-off in the theory. It is still too early to tell if these conjectures are actually true.

III. Quantization Procedure

The usual formulation for the equations of general relativity is in terms of action principle (discussed in Lecture IV)

$$s = \int R\sqrt{-g}\; d^4x \qquad (1)$$

However, as mentioned above, this formulation is not convenient for quantization. Rather we desir a Hamiltonian formulation. Therefore, one of the first problems in quantizing gravitational field is to formulate general relativity in a Hamiltonian form. That is, to construct a Hamiltonian for the theory, to find the canonical variables, and to apply the ordinary commutation relations to obtain Eigen solutions. However, to obtain a Hamiltonian formulation of the theory is a difficult task in itself because of the general covariance of the theory.

IV. The Hamiltonian Formulation of General Relativity

The Hamiltonian equations for any system described by the canonical variables q_i and p_i are of the following form:

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \quad \text{and} \quad \dot{p}_i = -\frac{\partial H}{\partial q_i} \tag{2}$$

where H the Hamiltonian, is a function of the q's and p's. Given q_i and p_i initially we can then find their first derivatives from equations (2) in terms of these initial values By successive differentiations we can find all higher derivatives in terms of them We can thus expand the solution $q_i(t)$ in a power series about t_o as

$$q_i(t) = q_{io} + \dot{q}_{io}t + .. = q_{io} + \frac{\partial H}{\partial p_i}{}_o t + ... \tag{3}$$

and

$$p_i(t) = p_{io} + \dot{p}_{io}t + .. = p_{io} - \frac{\partial H}{\partial q_i}{}_o t + .. \tag{4}$$

Thus, in a conventional Hamiltonian theory a knowledge of the initial q's and p's leads to a unique determination of their values at any future time. This situation, however cannot hold in general relativity as the following considerations will show. Let us suppose that, given the ten components of the metric and their first time derivatives initially the metric in the future is uniquely determined from the field equations We can picture the situation schematically in Figure 1 below. Here we plot the metric as a function of time. The abscissa schematically represents the functional space of the metric.

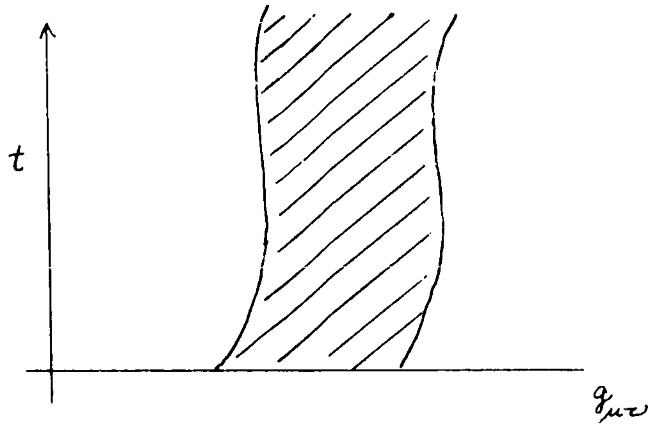


Figure 1 - Schematic representation of the time development of metric components.

However, we may perform a coordinate transformation which leaves everything unchanged up to some time $t_1 > t_o$ and thereafter deviates from the identity transformation. Such a transformation is a permissible transformation since all derivatives exist up to any order we desire. The effect of such a transformation is represented in Figure 2.

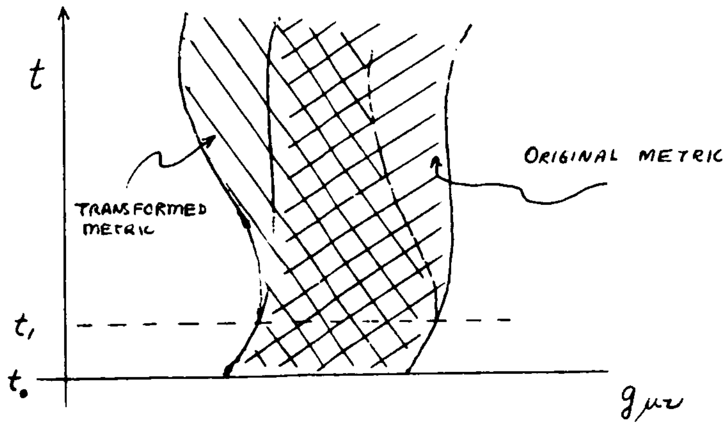


Figure 2 - Schematic representation of time development of metric components, under time dependent coordinate transformation.

In this way we obtain what appears to be a different solution to the field equations starting from the same initial data. Of course, the two solutions do not represent different physical situations but merely the same situation expressed in two different coordinate systems.

In Figure 2, the graph of the transformed metric is superimposed on the original metric. The region where the two graphs overlap represents that part of the metric which describes the physical situation and is not affected by a change of coordinates

From these considerations one can conclude that the field equations do not determine the time development of the metric uniquely. In the Hamiltonian formulation of the theory this non-uniqueness is reflected in a non-uniqueness in the Hamiltonian. If it were uniquely determined then of course one could obtain a unique solution for the metric in the manner indicated by equations (3) and (4).

There is another difficulty which arises in a Hamiltonian formulation of general relativity as a consequence of the general covariance of the theory. One can construct a momentum density $p^{\mu\nu}$ conjugate to $g_{\mu\nu}$ by differentiating the Lagrangian density of the theory with respect to $\dot{g}_{\mu\nu}$:

$$p^{\mu\nu} = \frac{\partial \mathcal{L}}{\partial \dot{g}_{\mu\nu}} \tag{5}$$

Actually the Lagrangian density $\mathcal{L}$ used here is not equal to $\sqrt{-g}R$ but differs from it by an ordinary divergence $(\varphi^{\mu}{}_{,\mu})$ so so chosen that the altered Lagrangian density is free of second derivatives of the metric. Both Lagrangian densities, of course, lead to the same field equations for the metric. Thus $p^{\mu\nu}$ is a well-defined functional of $g_{\mu\nu}$, $\dot{g}_{\mu\nu}$ and their spatial derivatives,

$$p^{\mu\nu} = p^{\mu\nu}(g_{\alpha\beta}, \dot{g}_{\alpha\beta}) . \tag{6}$$

However, because of the covariance of the theory it turns out that the canonical variables $g_{\mu\nu}$, $p^{\mu\nu}$ are not independent of each other but are related by a set of four equations of constraint, called the primary constraints, of the form

$$p^{0\mu} + \Gamma^{\mu}(g_{\alpha\beta}) = 0 \tag{7}$$

As a consequence we cannot invert equations (6) to obtain unique expressions for the $\dot{g}_{\mu\nu}$ in terms of the $p^{\mu\nu}$ and $g_{\mu\nu}$. Consequently, when one tries to construct the Hamiltonian density, $\mathcal{H}$, with

$$\mathcal{H} = p^{\mu\nu}\dot{g}_{\mu\nu} - \mathcal{L}(g_{\alpha\beta}, \dot{g}_{\alpha\beta}) \tag{8}$$

one cannot eliminate the $\dot{g}_{\alpha\beta}$'s from the right-hand side of eq. (8) to obtain $\mathcal{H}$ as a function of the canonical variables alone. This was one of the main problems that confronted people in formulating a Hamiltonian theory for general relativity. It was solved in different ways by Bergmann, Pirani and Schild[8] and Dirac[9].

A similar situation occurs in electrodyn ics. There things are simple enough so that one can see in detail just what is happening. The electromagnetic field is described by a vector potential $\underset{\sim}{A}$ and a scalar potential φ together with any matter variables such as irac fields which might occur. The theory is invariant under the group of gauge transformations

$$\underset{\sim}{\overline{A}} = \underset{\sim}{A} - \nabla\lambda \tag{9}$$

and

$$\overline{\varphi} = \varphi + \dot{\lambda} \tag{10}$$

where λ is an arbitrary space-time function. In general relativity the transformation group depends upon four arbitrary e-time functions. They are the four new coordinates expressed as functions of the old coordinates. In electrodynamics there is just one arbitrary function. However, many of the consequences are the same. Thus, by means of arguments similar to those used in the general relativity case one can show that a knowledge of the field quantities $\underset{\sim}{A}$ and φ together with their first time derivatives does not lead to a unique solution for these quantities into the future. Thus all of our comments concerning the Hamiltonian formulation of general relativity apply in this case with equal force.

The Lagrangian density of electrodynamics is given by

$$\mathcal{L} = \tfrac{1}{2}(\dot{\underset{\sim}{A}} + \nabla\varphi)^2 - \tfrac{1}{2}(\nabla \times \underset{\sim}{A})^2 - \rho\varphi + \underset{\sim}{j}\cdot\underset{\sim}{A}. \tag{11}$$

We can define the momentum density $\underset{\sim}{p}$ conjugate to $\underset{\sim}{A}$ as the derivative of $\mathcal{L}$ with respect to $\dot{\underset{\sim}{A}}$ and so obtain

$$\underset{\sim}{p} = \dot{\underset{\sim}{A}} + \nabla\varphi \qquad (12)$$

We note, however that $\mathcal{L}$ does not contain any $\dot{\varphi}$ terms so that π, the momentum conjugate to φ satisfies the equation

$$\pi = 0 \qquad (13)$$

This is the primary constraint associated with the gauge invariance of the theory and is analogous to the primary constraint equations (7). Here we see directly that we cannot determine $\dot{\varphi}$ in terms of the momentum densities However, we can obtain the $\dot{\underset{\sim}{A}}$'s in terms of the canonical variables and so obtain a Hamiltonian H given by

$$H = \int \mathcal{H}\, d^3x \qquad (14)$$

where the Hamiltonian density $\mathcal{H}$ is

$$\mathcal{H} = \tfrac{1}{2}\underset{\sim}{p}^2 + \tfrac{1}{2}(\nabla x \underset{\sim}{A})^2 - \underset{\sim}{j}\cdot\underset{\sim}{A} + \varphi(\nabla\cdot\underset{\sim}{p} + \rho) + \dot{\varphi}\pi \qquad (15)$$

In this expression, $\dot{\varphi}$ is taken to be an arbitrary space-time function. Its appearance reflects the non-uniqueness in the Hamiltonian which is necessary if the canonical equations of motion are not to determine the canonical variables uniquely in terms of their initial values

Unfortunately, this is not the whole story. There is another constraint equation that arises as a consequence of the requirement $\dot{\pi}$ be zero so that equation (13) is maintained throughout the evolution of the system. The time derivative of π is obtained in the usual way by taking its Poisson bracket with H. When we do this we find that

$$\dot{\pi} = (\pi, H) = \nabla \cdot \underset{\sim}{p} + \rho \qquad (16)$$

so that we must require that

$$\nabla \cdot \underset{\sim}{p} + \rho = 0 \qquad (17)$$

This is just one of Maxwell's equations since $\underset{\sim}{p}$ is equal to $-\underset{\sim}{E}$, the electric field Equation (17) is referred to as the secondary constraint of the theory. Fortunately all higher time derivatives of π and all time derivatives of $\nabla \cdot \underset{\sim}{p} + \rho$ vanish so that there are no additional constraints associated with the theory.

Obtaining a Hamiltonian formulation of general relativity was carried out along similar lines. However, because of the complexity of the primary constraints (7) the resulting expressions for the Hamiltonian and the secondary constraints were virtually impossible to work with. Recently Dirac, DeWitt, and myself, all independently, were able to introduce a new set of canonical variables into the theory in such a way that the new primary constraints took on the simple form

$$p^{o\mu} = 0. \qquad (18)$$

In terms of these new variables the Hamiltonian density took on the relatively simple form

$$\mathcal{H} = (g^{oo})^{-\frac{1}{2}}\mathcal{H}_L + g_{or}\mathcal{H}^r \qquad (19)$$

where $\mathcal{H}_L$ and $\mathcal{H}^r$ are certain functionals of the g_{rs} and p^{rs} and their spatial derivatives. (Here Latin indices take on the values 1, 2, 3,).

Since we require $p^{o\mu} = 0$ for all times, $\dot{p}^{o\mu}$ must also be zero for all times. $\dot{p}^{o\mu}$ is calculated by computing the Poisson bracket between $p^{o\mu}$ and the Hamiltonian H. Since $\mathcal{H}_L$ and $\mathcal{H}^r$ do not depend on $g_{o\mu}$ or $p^{o\mu}$, by taking the commutator of with respect to $p^{o\mu}$, one can further show that the secondary constraints become

$$\mathcal{H}_L = \mathcal{H}^r = 0 \tag{20}$$

The constraint on $\mathcal{H}^r$ is known as the longitudinal covariance, that on $\mathcal{H}_L$ is known as the Hamiltonian constraint. These constraint equations are the main cause of all difficulties in formulating a quantized version of the theory. The existence of these constraints is a direct consequence of invariance of the theory under arbitrary coordinate transformations. For this reason it is most likely that the difficulties associated with the Hamiltonian formulation of the theory will generally appear in one way or another in any formulation of the theory. In the present formulation, they tell us that the canonical variables g_{rs} and p^{rs} are not independent of each other. But in formulating a Hamiltonian quantization by imposing commutation relations on canonical variables it is essential that these variables be independent. This means that, in effect, we have too many variables and some should be eliminated from the theory. Unfortunately, the standard methods of eliminating the redundant parts of $g_{o\mu}$ and $p^{o\mu}$ are not directly applicable because of the complexity of the constraint equations (20).

A simplified form for these constraints was first given by Dirac. The equation for $\mathcal{H}^r$ reduces to

$$\mathcal{H}^r = p^{sr}{}_{|s} \tag{21}$$

Subscript $|s$ denotes covariant differentiation using the metric g_{rs} and its inverse e^{rs}. It is not the full four-dimensional covariant derivative, but only the three-dimensional covariant derivative.

Equation (21) is very similar to the equation that appears in electrodynamics in the case of zero charge density.

$$\nabla \cdot p = 0 \tag{22}$$

Equation (21) is a generalization of the divergence applied to a symmetric tensor in curved spaces. In electrodynamics we have simply the ordinary divergence of a vector. But this difference is the cause of many difficulties. The constraint on $\mathcal{H}_L$ is

$$\mathcal{H}_L = \frac{1}{K}\left(g_{ra}g_{sb} - \tfrac{1}{2}g_{rs}g_{ab}\right) p^{rs}p^{ab} + R(g_{ab}) \tag{23}$$

where

$$K^2 = |\, g_{rs} \,| \tag{24}$$

and ${}^3R(g_{ab})$ is the curvature scalar constructed from the metric g_{rs} and its inverse. The first term resembles a kinetic energy while the second term resembles potential energy. In the linearized version of the theory these terms are in fact interpreted as kinetic and potential energy.

In order to understand better the type of difficulties introduced into the theory by the constraint equations (20), let us return to our example of electrodynamics. There we have the variables $\underset{\sim}{A}$ and $\underset{\sim}{P}$. They are not independent variables, because they satisfy Equation (18). This means that not all of the variables of the theory are independent dynamical variables. If we can perform some kind of transformation and make $\nabla \cdot p + \rho$ a new momentum density for the theory, then together with the canonical coordinates conjugate to the new momentum they will play the same role as π and Φ and can be eliminated from the theory.

One very simple way of doing this is to introduce the longitudinal and transverse components of $\mathbf{A}$ and $\mathbf{P}$.

Let

$$\mathbf{A} = \mathbf{A}^L + \mathbf{A}^T \tag{25}$$

where $\mathbf{A}^L$ and $\mathbf{A}^T$ satisfy the following conditions:

$$\nabla \cdot \mathbf{A}^T \equiv 0 \tag{26}$$

$$\nabla \cdot \mathbf{A}^L \equiv 0 \tag{27}$$

Similarly, $\mathbf{P}$ may be written as

$$\mathbf{P} = \mathbf{P}^L + \mathbf{P}^T$$

Then Equation (18) reduces to

$$\nabla \cdot \mathbf{P}^L + \rho = 0 \tag{28}$$

$\mathbf{P}^T$ does not appear in the constraint equation and we are free to consider $\mathbf{P}^T$ and $\mathbf{A}^T$ as the basic dynamical variables. $\mathbf{P}^L$ is expressed in terms of ρ. If ρ is zero $\mathbf{P}^L$ is also zero. $\mathbf{A}^L$ may be made to be zero by introducing a proper gauge condition. Hence one can construct the Hamiltonian in terms of $\mathbf{A}^T$ and $\mathbf{P}^T$

which are then the canonical variables.

The commutation relations between $\underset{\sim}{A}^T$ and $\underset{\sim}{P}^T$ contain some terms other than the usual delta function commutation relations. However, these terms are independent of $\underset{\sim}{A}^T$ and $\underset{\sim}{P}^T$ and no new complications arise. $\underset{\sim}{A}^T$ and $\underset{\sim}{P}^T$ are also invariant under a gauge transformation. Under a gauge transformation only $\underset{\sim}{A}^L$ changes. In Figure 3 we have schematically represented two different A fields describing the same physical situation, i.e., the same E and B fields.

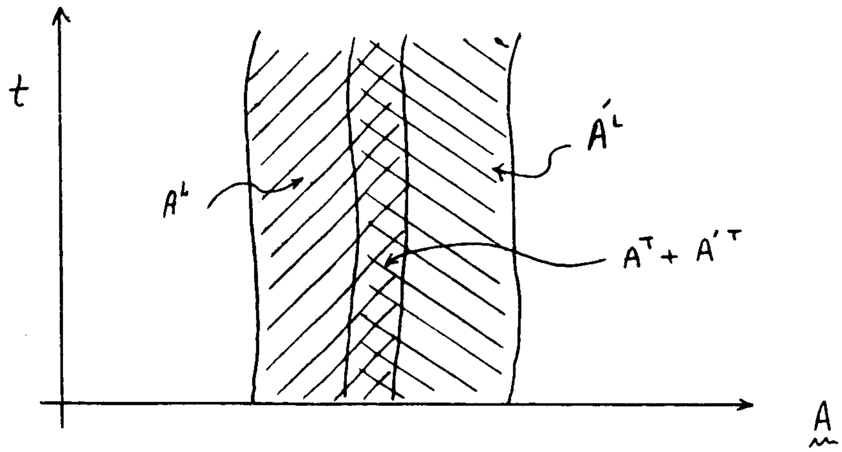


Figure 3 - Schematic representation of A field under gauge transformation.

The central portion of the figure represents the transverse parts of the two A fields and is the same for both. The two outer portions represent the different longitudinal parts of the two fields.

With almost no modification, the above result applies to the case of gravitation. That is, in the p^{rs} - g_{rs} representation, there is some invariance under coordinate transformations. This represents the intrinsic geometry. In figure 4, the non-variant part of the representation is indicated by the fluff about an invariant core. This fluff depends upon the particular coordinate system one chooses. Under any coordinate transformation, the central core remains unchanged.

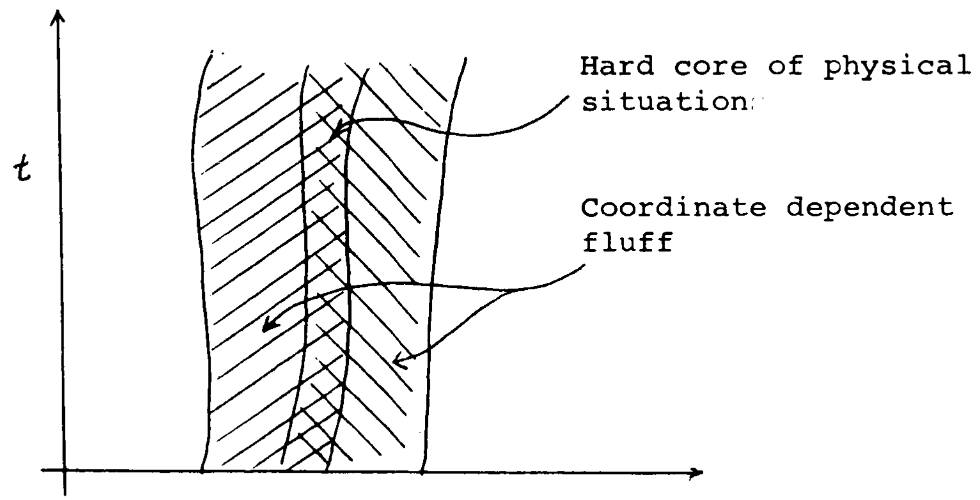


Figure 4 - Schematic representation of invariant core of intrinsic geometry and coordinate dependent fluff.

As in the electromagnetic case, one would like to separate from g_{rs} and p^{rs} a physical part which remains unchanged under coordinate transformations. How easily this separation can be made depends upon the form of the constraint equations. In the electromagnetic case, the separation is achieved by breaking up $\underset{\sim}{A}$ and $\underset{\sim}{P}$ into longitudinal and transverse parts. As we saw, the longitudinal part of p is uniquely determined by the constraint equation (28). On the other hand, A^L can be transformed away by choosing a proper gauge condition.

It is possible to break up the g_{rs} and p^{rs} in a manner analogous to the electromagnetic example. However, the redundant variables cannot be eliminated from the theory very easily.

At best, one can accomplish this solution by an approximation procedure based on a weak field approximation. To date, no one, to my knowledge, has suggested a decomposition scheme of g_{rs} and p^{rs} such that the physical part does not appear in the constraint equations, or a scheme allowing one to solve the constraint equations directly and to write down the redundant variables in terms of the other variables.

V. Quantum Version of the Theory

With the above discussion of the classical Hamiltonian formulation of general relativity and electrodynamics we can now turn our attention to the quantum versions of these theories. Again, many of my remarks will be devoted to the electromagnetic case since we know fairly well what is going on there. There are two different ways to quantize these theories within the Hamiltonian framework. The most straightforward of these, and the one used by Bergmann and Komar, and DeWitt,[10] is to treat only the physical part of the field varaibles as operators defined in some Hilbert space. The remaining field variables are to be eliminated from the theory by the use of the constraint equations and the imposition of gauge or coordinate conditions. Thus, in the electromagnetic case, one would treat only $\mathbf{A}^T$ and $\mathbf{p}^T$ as operators and replace $\mathbf{P}^L$ by $-\nabla \frac{1}{\nabla^2}\rho$. Once one has fixed the gauge (for example, that $\mathbf{A}^L = 0$) one can obtain the Hamiltonian directly in terms of these transverse parts. Then one can write down the Schrödinger equation. Finally the commutation relations between the transverse parts follow from their Poisson bracket relations in the usual way. A possible representation analogous to the x-representation in ordinary quantum mechanics would be to define the operators $\mathbf{A}^T$ and $\mathbf{P}^T$ as follows:

$$
\begin{aligned}
\mathbf{A}^T\varphi &= \mathbf{A}^T\varphi \\
\mathbf{P}^T\varphi &= i\hbar\frac{\delta}{\delta\mathbf{A}^T}\varphi
\end{aligned}
\tag{29}
$$

In this way, state vectors would then be functionals of $\mathbf{A}^T$.

We may apply the same procedure to the gravitational case for which g_{rs}^{TT} and p^{rsTT} correspond to the transverse parts of the canonical coordinates and momenta. Using a representation similar to equation (29) we take

$$
\begin{aligned}
g_{rs}^{TT}\varphi &\equiv g_{rs}^{TT}\varphi \\
p^{rsTT}\varphi &\equiv i\hbar \frac{\delta\varphi}{\delta g_{rs}^{TT}}
\end{aligned}
\tag{30}
$$

State vectors would then be functionals of g_{rs}^{TT}. However, because we could not solve the constraint equations for any four of the redundant variables in closed form, we do not have a closed form for the Hamiltonian in terms of the g_{rs}^{TT} and p^{rsTT}. We are forced to make a weak field expansion. This is a return to the linearized theory which Feynman has so nicely treated.

There is another difficulty which arises when one takes this approach to quantization. The above scheme for separating off the physical part is by no means unique in either the electromagnetic or the gravitational case. One can set up a scheme which allows one to calculate the physical part in many different ways. For example, in electrodynamics one can fix the gauge by imposing conditions on some of the $\underset{\sim}{A}$'s. The remaining $\underset{\sim}{A}$'s will then be gauge invariant. As an example, one can fix the gauge by requiring that

$$\mathbf{A}_1 = 0 \tag{31}$$

$$\mathbf{A}_2 \ (x = 0) = 0 \tag{32}$$

$$\mathbf{A}_3 \ (x = y = 0) = 0 \tag{33}$$

In this gauge, the values of $\mathbf{A}_2$ $(x \neq 0)$ and $\mathbf{A}_3$ $(x \neq 0, y \neq 0)$ are gauge invariant quantities and together with their canonical conjugate momenta can be used to describe the physical state of the electromagnetic field. This description is, of course, quite different from the previously discussed condition $\mathbf{A}^L = 0$. Thus, there are many different ways of fixing the gauge in electrodynamics. Each different way in turn leads to a different set of expressions which can serve as the physical variables.

The situation in general relativity is quite analogous. Once the coordinates are fixed by imposing suitable conditions on the g_{rs}'s and p^{rs}'s, the remainder variables automatically become invariant under coordinate transformations and can serve as physical variables. These are commonly referred to as observables. An interesting question is whether a quantized version using one set of observables is equivalent to a quantized version using another set. I will discuss some of the difficulties involved in answering that question.

Imagine that two different decompositions of the g_{rs} and p^{rs} have been found. Symbolically we write

$$\{g_{rs}, p^{rs}\} = \{y_{physical}, y_{coordinate}\} \tag{34}$$

and

$$\{g_{rs}, p^{rs}\} = \{y'_{physical}, y'_{coordinate}\} \tag{35}$$

The y's represesent physical or coordinate conditions. Since $y_{physical}$ and $y'_{physical}$ both represent the same physical conditions, they must be functions of each other. However, in general, $y'_{coordinate}$ will depend both on $y_{coordinate}$ and $y_{physical}$ and vice versa. As for example in the formulation of Equation (21), we wish to treat only the physical parts of the field as operators and to consider the coordinate part as a c-number (classical, or commuting number, that is, not an operator). This separate treatment of the physical and coordinate parts leads to the following difficulty: what is treated as a c-number in one formulation will appear as an operator in the other formulation, and vice versa. It is not inconceivable that one can eliminate the above-mentioned difficulties and construct a general proof of the physical equivalence of different decomposition schemes, which are within the framework of the quantization procedure discussed above. However, I strongly doubt this possibility for reasons which I will now discuss.

We have mentioned before that there are actually two ways of affecting a quantization within the Hamiltonian framework. In the one, discussed above, only the physical parts of the field are to be treated as operators defined in a Hilbert space.

The physical parts are invariant under a gauge or coordinate transformation. Hence the gauge or coordinate group disappears from this "Hilbert space quantization," making it difficult to carry out the proof of equivalence discussed above.

In the other approach to quantization, one treats all of the field variables as operators. These operators, either the $\underset{\sim}{A}$ and $\underset{\sim}{P}$ or the g_{rs} and p^{rs}, are assumed to have the standard commutation relations between canonical variables. Thus we have, for example

$$[g_{rs}, g'_{ab}] = [p^{rs}, p'^{ab}] = 0 \tag{36}$$

$$[g_{rs}, p'^{ab}] = i\hbar\delta^{ab}_{rs}\delta(x - x') \tag{37}$$

The prime over a variable means that it is evaluated at the space point x'. These operators operate in a linear vector space whose elements are functionals of the g_{rs}. In this representation the operators are given by

$$\begin{aligned} g_{rs}\varphi(x) &= g_{rs}\varphi(x) \\ p^{rs} &= i\hbar \frac{\delta}{\delta g_{rs}} \varphi(x) \end{aligned} \tag{38}$$

In this case all the coordinates and momenta and not just the transverse (physical) parts are treated as operators. Because of the constraint equations (20), this vector space is not normalizable and hence is not a Hilbert space.

We now ask, how do these constraints modify the quantization of the theory. With the Hilbert space method of quantization, the constraints were no problem because they were eliminated from the theory prior to quantization. Now they must be taken into account. However, the constraint equations cannot be treated simply as operator equations directly. If we do this, they will be inconsistent with the commutation relations (38). That is, if we evaluate the covariant derivative of both sides of equation (37) at x'^b and if we assume that $\mathcal{H}'^a = p'^{ab}{}_{|b} = 0$ everywhere, the left hand side will be zero everywhere, while the right hand side will not be zero at many points of space. There is one way to avoid this difficulty. To describe the physical states of the gravitational field, we shall use only those elements of the linear vector space, ψ, which satisfy

$$\mathcal{H}^a \psi = 0 \tag{39}$$

and

$$\mathcal{H}_L \psi = 0 \tag{40}$$

It is possible to show that, with these restrictions on the state vectors, the quantum version of the theory passes over to the classical version in the correspondence limit. This will not be so if this assumption is not imposed.

In order to complete this formulation we must impose addition gauge or coordinate conditions. However, as is the case of the constraint equations, they cannot be treated as operator equations but must be assumed to hold only for a subclass of the state vectors which satisfy conditions (25a) and (25b). Thus the subspace of the original linear vector space spanned by the vectors which satisfy these conditions is further broken up into subspaces by the imposition of various kinds of gauge or coordinate conditions. We schematically picture the space of all vectors defined by equation (24), in Figure 5. The shaded area represents the subspace of vector space of vectors satisfying conditions (25a) and (25b). The two smaller cross hatched areas represent subspaces, in which two different sets of coordinate conditions hold.

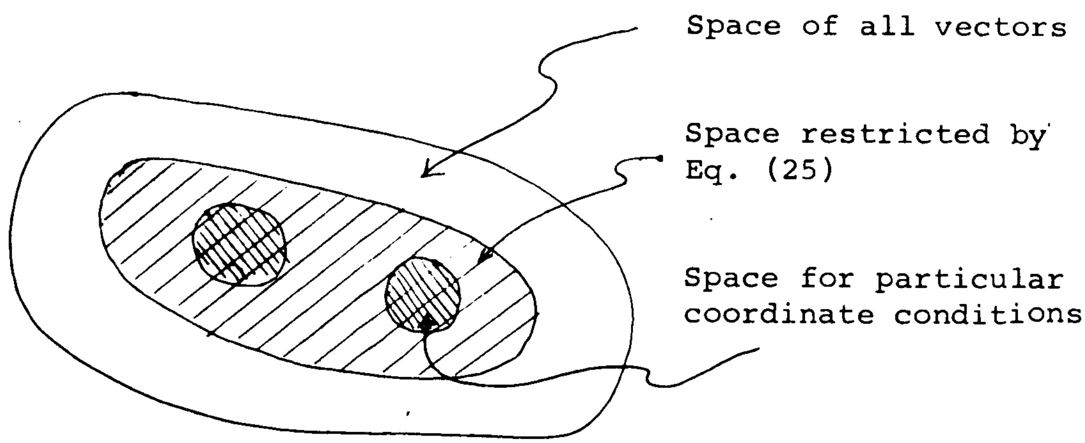


Figure 5 - Schematic representation of vector space

We can draw a similar picture in the classical theory as well, only now the overall space is the phase space with coordinates $\mathbf{A}$ and $\mathbf{P}$ or g_{rs} and p^{rs}. The shaded area now will represent the subspace of points satisfying the constraint equations while the two smaller, cross hatched areas will represent points satisfying two different sets of gauge or coordinate conditions. As we have seen, the points in the large shaded area are, in a certain sense, redundant; many of them represent the same physical situation. They differ only because different coordinates are used to describe the same state of the field. Thus we should expect that the totality of points within any one cross hatched area should stand in a one-to-one relation with the possible states of the system. This also means that there should be a one-to-one correspondence between the points of one cross-hatched area and those of any other cross hatched area.

Indeed, it has been proved in the classical version of the theory that there always exists a canonical transformation which maps the points of one cross hatched area in a one-to-one manner onto the points of any other cross hatched area.[11]

These transformations are generated by linear combinations of the constraints. Since the transformation, which maps one cross hatched area onto another, is obviously a finite gauge or coordinate transformation, we see that linear combinations of the constraints form the generators of the invariance group of the theory. In the course of the proof it was necessary to show that these generators did indeed form a group. The necessary and sufficient condition for this to be true is that the Poisson bracket between any two constraints is a linear combination of constraints. The constraints do possess this property. It is thus concluded that the subspace spanned by those points in phase space, for which the constraint equations are valid (the shaded area in Figure 5), is simply connected. Therefore, there can be no physical experiment within the theory which could single out a preferred coordinate system. Thus the principle of general covariance is not violated.

From the above discussion, we can see the intimate relation between the invariance group of the theory and the constraint equations. This relationship is a general one and holds whenever a theory is derivable from a variational principle, and possesses an invariance group whose elements are defined by one or more space-time functions.

These concepts can be best illustrated in terms of electromagnetic theory. If we include φ and π, as well as $\underset{\sim}{A}$ and $\underset{\sim}{P}$ as our canonical variables, we can form the following generator of an infinitesimal gauge transformation:

$$c = \int d^3x \, \{\dot{\gamma}\pi + \gamma(\nabla \cdot \underset{\sim}{P} + \rho)\} \tag{41}$$

The variation of the field quantities φ and $\underset{\sim}{A}$ may be easily computed in the subspace of interest in which the constraint equations are also satisfied:

$$\bar{\delta}\Phi = \{\Phi, c\} = \dot{\gamma} \tag{42}$$

and

$$\bar{\delta}\underset{\sim}{A} = \{\underset{\sim}{A}, c\} = -\nabla\gamma \tag{43}$$

Equations (42) and (43) represent an infinitesimal gauge transformation. Also, it is easy to prove that the generators π and $\nabla\cdot\underset{\sim}{P} + \rho$ have vanishing Poisson brackets among themselves so that the generators of the form given in equation (41) do indeed form an infinitesimal group. The elements of this infinitesimal group can be added to give a finite gauge transformation. Exactly the same situation pertains in general relativity where a linear combination of the constraints generate an infinitesimal coordinate transformation.

Let us now return to the quantum version of electromagnetic theory. What we have said about the classical theory is also valid in the quantum theory if we replace canonical transformations by unitary transformations and Poisson brackets by commutators. In the electromagnetic theory this is easily done; when the canonical variables are treated as operators, c is the generator of an infinitesimal unitary transformation. The commutator of two such transformations is a unitary transformation of the same kind.

It is not difficult to show that the two methods of quantization outlined above are equivalent. We may start from the classical theory, impose gauge conditions and then quantize, treating only the physical parts of the field as operators in a Hilbert space; or we may treat all of the field variables as operators in a linear vector space and impose gauge conditions afterwards to restrict the vectors which are used to describe the physical state of the system. The essential point of the latter formulation is, that we can carry out a unitary transformation which transforms from one coordinate frame to another. From this we can prove the equivalence of all gauge frames. Since the two methods are equivalent, we have thereby proved the equivalence of starting out in two different classical gauge frames.

There is another difficulty which appears in equations (42, 43) to be overcome. In order to generate a transformation from one set of potentials to another which satisfies a particular set of gauge conditions, the gauge function, γ, will in general depend upon the potentials. Thus, if we wish to transform an arbitrary set of potentials to the Coulomb gauge where $\nabla \cdot \underset{\sim}{A} = 0$, we must use a gauge function given by

$$\Gamma = \frac{1}{\nabla^2}\nabla\cdot\bar{\underset{\sim}{A}} \tag{44}$$

where $\bar{\underset{\sim}{A}}$ is the original potential. This has the consequence, that if $\bar{\underset{\sim}{A}}$ and $\bar{\underset{\sim}{P}}$ are assumed to satisfy the standard commutation relations

$$[\bar{A}_r, \bar{P}'^s] = i\hbar\delta_r^s\delta \tag{45}$$

then it follows that

$$\underset{\sim}{A} = \bar{\underset{\sim}{A}} - \nabla\frac{1}{\nabla^2}\nabla\cdot\bar{\underset{\sim}{A}} \tag{46}$$

and

$$\underset{\sim}{P} = \bar{\underset{\sim}{P}} \tag{47}$$

will not satisfy equation (45). To by-pass this difficulty, one makes use of equations (42,43) for generating gauge transformations. Since C is the generator of an infinitesimal unitary transformation the transformed variables will satisfy the commutation relations (45). Equations (42, 43) are valid only in the subspace of the linear vector space where the constraint equations hold. In the full linear vector space there are additional terms in the expressions for $\delta\bar{\psi}$, $\delta\bar{\underset{\sim}{A}}$, $\delta\bar{\pi}$ and $\delta\bar{\underset{\sim}{P}}$ which are linear combinations of the constraints. These terms arise from the non-vanishing commutators between ψ, π, $\underset{\sim}{A}$ and $\underset{\sim}{P}$ and the quantities $\dot{\gamma}$ and γ. They maintain the validity of the commutation relations between the transformed field variables. It is then a straightforward procedure to prove the invariance of the theory under arbitrary q-number (non-commuting quantum operator) gauge transformations even in the presence of interaction with a Dirac field.

We can illustrate these remarks with a simple example (fig. 6). Two observers, A and B, observe the electrical properties of a resistor. They both realize that the potential difference between two points is the only meaningful concept, but that it is much easier to assign an arbitrary value of potential to one fixed point and then measure all other potentials with respect to that fixed point.

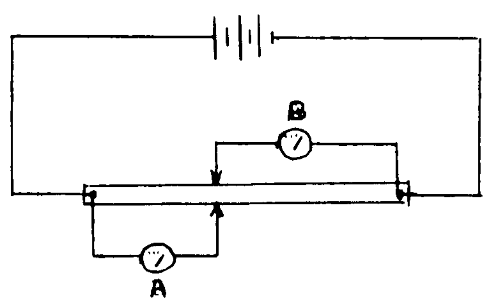


Figure 6 - Illustration of effect of fluctuating gauge transformation by measurement of potential difference along a resistor.

A decides to assign the value zero to the potential at the left end of the resistor, while B decides to assign the value zero for the potential to the right side of the resistor. For an idealized resistor they will have no trouble comparing their results. For instance, B will merely add to all of his potential readings the value of the total potential difference between the two ends of the resistor. In order to transform his result to that of A's, B performs a gauge transformation. Their results are then directly comparable.

Now let the resistor cease to be ideal. They will be looking at a resistor in which the potential is fluctuating due to thermal noise. A will conclude that the potential at the right end fluctuates, while B will conclude that the potential at the left end fluctuates. This will puzzle both of them for a moment since each had assigned the value zero to the potential at one of these two ends respectively, and by definition the value zero does not fluctuate. In order for B to check his results with those of A, he will now have to perform a gauge transformation which fluctuates with time. Such a transformation is the analogue of a q-number gauge transformation in the quantum theory.

In general relativity the situation is much the same as in electrodynamics with one crucial difference. While one can still construct the generator of an infinitesimal q-number coordinate transformation, one cannot integrate to obtain a finite transformation. The infinitesimal generators do not appear to form a group. In fact, one can prove that there exists no ordering of factors in the classical expressions for the constraints such that the commutator of ${}_L$ and ${}'_L$ is again a linear combination of the constraints. Thus it appears that the shaded area in Fig. 5 is not simply connected; the use of different coordinate conditions appears to lead to essentially non equivalent quantum theories. This conclusion stands in direct contradiction with the principle of general covariance. If true, it would mean that, in principle at least, it should be possible to decide which, of the infinity of possible coordinate systems, is the one appropriate for a description of our universe.

V. Conceptual Problems in Quantized General Relativity

The field variables of general relativity, the $g_{\mu\nu}$, play a dual role in the theory. On the one hand they describe the gravitational field while on the other they serve as a metric and so determine the geometry of space-time and hence affect all other fields that exist in the space-time manifold. If we now consider the $g_{\mu\nu}$ as quantum field variables they will exhibit the customary quantum fluctuations. As long as we think of the $g_{\mu\nu}$ as describing the gravitational field this additional complication seems to offer no more difficulty in understanding than in the electromagnetic case when we quantize that field. However, when we also use the $g_{\mu\nu}$ to describe the metric, many new conceptual problems arise not the least of which is what do we mean by a fluctuating geometry. I will conclude this lecture with a brief discussion of some of these problems.

Many of the conceptual problems associated with quantizing general relativity are related to measurement processes. Since general relativity is first of all a geometrical theory, the most natural types of measurements will involve the determination of space-time intervals, measurement of the gravitational field $g_{\mu\nu}$ and the Christoffel symbols $\left\{ {\rho \atop \mu\nu} \right\}$ in a given coordinate system. Wigner[12] has investigated the question of measuring time intervals in general relativity. He concluded that there seems to be a contradiction.

Wigner's argument is as follows: When one measures an interval of time in some region of space, what one actually measures is how long it takes for something to happen. For this purpose one needs an accurate clock. The accuracy in time measurement is limited by the quantum uncertainty relation

$$\Delta E \Delta t \geq h \tag{48}$$

Therefore, if one wants an accurate clock (small Δt) then there will be an uncertainty in its energy, and this uncertainty in its energy is related to an uncertainty in its mass through the relation

$$(\Delta m)c^2 = \Delta E \tag{49}$$

Thus the smaller Δt is, the bigger Δm must be. This means the fluctuations in the clock mass must be very large. In the limit of infinite accuracy the fluctuations in the clock mass must be infinite. These fluctuations in mass will in turn produce similar fluctuations in the gravitational field. In special relativity, where the metric is forever fixed, such fluctuations are not a problem since one need not include the gravitational field in the theory. In general relativity we can no longer neglect such fluctuations since the gravitational field is also the metric field and hence affects all other fields. As a consequence the very notion of a space-time interval and with it the notion of a point in space-time become questionable and the whole nature of the space-time manifold uncertain. It has even been conjectured, on the basis of such arguments, that one must eliminate the concept of points from the theory since they are unobservable elements in the theory. At the very least it does bring into question the process of setting up a coordinate system and the measurement of distances

A possible way out of this apparent dilemma is to abandon the demand for accurate time measurement at a particular space point and adopt the approach of S-matric theory wherein one talks only of asymptotic behavior of interaction systems. Such a procedure has been suggested by Misner. However, I believe we can learn about the structure of space and time only by examining the character of the coordinate system

This does not mean that we have to actually observe these coordinates. Rather our structure of space-time will determine to a large extent the type of coordinates we can introduce For example, if space-time were in some way discrete then we would employ different coordinates from those we would employ if it were continuous. This in turn would affect the types of theories we could construct in space-time. Thus, even if we were never able to make local space-time measurement of the kind envisaged by Wigner, we could still have a space-time manifold whose properties would be known to us through the properties of the coordinate systems that exist in it.

Another difficulty arises when we try to work out a Bohr-Rosenfeld argument for the measurement of a gravitational field.[13] Bohr and Rosenfeld noted that in electromagnetic theory, when one wants to measure the electric field with great accuracy, one has to employ very large charges. Furthermore, since the measurement of the field involves a measurement of the momentum imparted to the particle we must, in some way minimize the uncertainty in this measurement. The uncertainty, ΔE_x, in the x component of the electric field is related to the uncertainty ΔP_x, of the x component of the momentum of the test particle by

$$\Delta E_x = \frac{\Delta P_x}{\epsilon \Delta t} \tag{50}$$

where ϵ is the charge on the test body and Δt is the time taken for the measurement of E_x. The uncertainty in the momentum measurement is itself limited by the uncertainty relation $\Delta P_x \Delta x \geq \hbar$ so that we have

$$\Delta E_x \geq \frac{\hbar}{\epsilon \Delta x \Delta t} \tag{51}$$

where Δx is the uncertainty in the position of the test particle. Since we want all three quantities, ΔE_x, Δx and Δt to be small, we must use a test body with a very large charge. However, a large ϵ entails a further difficulty since we measure the total field present. This includes the field produced by the test-body. Thus, for large ϵ, this self-field, $\mathcal{E}$, is also large. Furthermore, we obtain no knowledge of this field since we do not know the position and motion of the test body exactly because of the uncertainty relations. However, Bohr and Rosenfeld showed that, by the use of purely mechanical devices such as springs, it is possible to compensate automatically for the effect of this self-field. They concluded that it is meaningful to talk about the value of an electric field at a point in space-time.

In the gravitational case one can write an expression analogous to Eq. (35). Here the objects of measurement are components of the Christoffel symbols $\left\{ {\rho \atop \mu\nu} \right\}$. The component analogous to E_x is $\left\{ {1 \atop 44} \right\}$. If we assume that the test body moves along a geodesic then we can show, in a manner analogous to the derivation leading to Eq. (35), that the uncertainty in $\left\{ {1 \atop 44} \right\}$ is:

$$\Delta\left\{ {1 \atop 44} \right\} \geq \frac{\hbar}{m\Delta x \Delta t} \tag{52}$$

where m is the gravitational mass of the test body. Again, in order to specify the graviational field at a space-time point, we must make all three quantities, $\Delta\left\{ {1 \atop 44} \right\}$, Δx, and Δt vanishingly small. The requirement that $\Delta\left\{ {1 \atop 44} \right\}$ is small means that m must be very large. This large mass will produce a large field. But now the effects of this field cannot be compensated for as in the electromagnetic case. First of all, there are no devices such as springs which will not produce a gravitational field. Second, we have no way of determining the amount of compensation needed. Third, once the mass of the test body becomes large, it will no longer follow along a geodesic of the field we are trying to measure, but will follow a geodesic of the combined fields. Fourth, since the theory is non-linear, the two fields, external and self-, will not add, so that we can no longer make use of the geodesic equation to measure the $\left\{ {\rho \atop \mu\nu} \right\}$.

In both of the above examples we are confronted with a new situation which does not occur in ordinary Lorentz invariant theories. Whenever we try to make a measurement of gravitational fields at a point we must use very heavy bodies to overcome the effects of the uncertainty principle. The introduction of heavy bodies as measuring instruments distorts the result in an unpredictable way.

Another difficulty is associated with setting up the initial value problem. The initial conditions are imposed upon space-like surfaces However, one must know the geometry of the space-time in order to pick out a space-like surface Classically, one can overcome this difficulty by choosing some surface determined by a condition on the coordinates such as t = 0, such that this surface is indeed space-like. For example, one can fix g_{rs} in a coordinate system where $g_{o\mu} = - \delta_{o\mu}$, such that, on the surface t = 0 $ds^2 = g_{rs}dx^r dx^s - dt^2 > 0$. Lichnerowicz and Mme. Foures-Bruhat showed that as the system develops, the adjoining surface, t = dt, is also space-like.

When we come to the quantum version of the theory, we are faced with the same problem. The initial state is represented by functions of g_{rs}. We can employ a coordinate system in which $g_{o\mu} = -\delta_{o\mu}$, and choose the initial surface to be that defined by t = 0. Then we must choose the ensemble of gravitational fields which makes up the eigen states of the field so that $ds^2 > 0$. This is easy although it imposes rather complicated restrictions on our initial state functionals Now, the problem is, can we be sure that our initial space-like surfaces remain space-like into the future for all numbers of the ensemble? In the classical case, for a given set of g_{rs}, one can find a corresponding set of momenta p^{rs} which satisfy the constraint equations which lead to the surface $t = dt$ being space-like. In the quantum version it is possible that a surface that was initially space-like might become time-like or light-like. Such a possibility exists because of the difficulty mentioned earlier that one cannot find quantum expressions for the constraints (20) such that the commutator of any two of them is a linear combination of the constraints with all coefficients standing to the left of the constraints. Since the Hamiltonian is itself a linear combination of the constraints we find that they are not necessarily zero for the physical states ψ which satisfy eqs (39) and (40). As a consequence the p^{rs} may have fluctuations that do not satisfy the constraint equations at later times and so lead to non-space-like surfaces.

Thus, both from the formal and the conceptual points of view, there are serious problems associated with quantizing gravity. In all our examples there appear to be incompatibilities between the requirements of quantum mechanics as we know them and the requirements of general relativity. It is of course possible that in a complete theory of quantum gravidynamics such difficulties might, in some way, be ameliorated. One can argue that it is meaningful to discuss the measurement process only on the basis of the complete theory. However, one does not yet have such a complete theory. Hopefully such discussions as we have given here, even with the crude methods used, might indicate this theory and the direction it should go, or whether it is reasonable to try to construct.

REFERENCES

1) P.G. Bergman, Phys. Rev. 75,680 (1949)

2) For a detailed bibliography the reader is referred to: P.G. Bergman, Helvetica Physica Acta, Suppl. IV, 79 (1956) and *Recent Developments in General Relativity*, New York, Pergamon Press, 31 (1962)

3) P.G. Bergmann and A Komar, *Les Theories Relativistes de la Gravitation*, Paris, Centre National de la Recherche Scientifique, 309 (1962)

4) B.S. DeWitt, *Recent Developments in General Relativity*, 175 (1962)

5) R.E. Peierls, Proc. Roy. Soc. (London), A 214, 143 (1952)

6) C.W. Misner, Rev. Mod. Phys. 29, 377 (1957)

7) R. Arnowitt and S. Deser, Phys. Rev. 113, 745 (1959)

8) F.A.E. Pirani and A. Schild, Phys. Rev., 79, 986 (1950)

9) P.A.M. Dirac, Proc. Roy. Soc. (London), A246, 333 (1958); Phys. Rev. 114, 924 (1959)

10) R. Arnowitt, S. Deser and C.W. Misner, *Recent Developments in General Relativity*, 127. Further references to their work are contained here.

11) J.L. Anderson and P.G. Bergmann, Phys. Rev. 83, 1018 (1951)

12) E.P. Wigner

13) J. L. Anderson, Rev. Mex de Fisica

MACH'S PRINCIPLE AS BOUNDARY CONDITION

FOR EINSTEIN'S FIELD EQUATIONS

Lecture XV

J. A. Wheeler

Princeton University

Seminar on Gravitation and Relativity, NASA Goddard Space Flight Center, Institute for Space Studies, New York, N. Y.; edited by H. Y. Chiu and W. F. Hoffmann.

Preface

These notes were taken from a series of seminars on Gravitation and Relativity presented at the Institute for Space Studies, NASA Goddard Space Flight Center during the academic year 1961-1962. Professor R. H. Dicke of Princeton University organized the series as an introduction to the subject for non-experts, emphasizing the observable implications of the theory and the potential contribution of space sciences may make towards a better understanding of general relativity.

The approach has been conceptual rather than formal. For this reason, this record does not include a complete mathematical development of the subject, but, we hope, does contain sufficient mathematics to elaborate on the conceptual discussions.

The notes were prepared with a certain amount of editing from a transcript made from recordings of the lectures. The speakers have not had the opportunity to read and correct the final manuscript. Hence, we accept responsibility for errors and omissions.

H. Y. Chiu

W. F. Hoffmann

I. THE SEARCH FOR AN ACCEPTABLE FORMULATION OF MACH'S PRINCIPLE

Inertia as a Consequence of an Interaction between the Accelerated Test Particle and all the Rest of the Universe

Acceleration can have no meaning unless there is something with respect to which the acceleration takes place. The acceleration with respect to absolute space that Newton speaks about has to be understood as acceleration with respect to the stars and matter in the universe. These two sentences state in oversimplified form the argument of Mach[(1)]. From it he went on to make conclusions about the origin of inertia. Inertia - being tied to acceleration - must arise from interaction between the object under study and all the other mass in the universe. Thus Mach's principle might be stated in this form: (Formulation 1). The inertial properties of an object are determined by the distribution of mass-energy throughout all space.

Inertia as the Radiative Component of the Gravitational Force

Mach's principle, together with Riemann's idea that the geometry of space responds to physics and participates in

---

(1) Ernst Mach, Die Mechanik in ihrer Entwicklung (Leipzig, 1st ed. 1883,..., 7th ed. 1912, 9th ed. 1933; latter translated into English by T. J. McCormack as The science of mechanics, Open Court Publishing Co., La Salle, Illinois, 1960).

physics, were the two great currents of thought which Einstein, through his powerful equivalence principle, brought together into the present day geometrical description of gravitation and motion. In the course of this work Einstein identified gravitation itself as the source of the interaction by which - according to Mach - one object affects the inertial properties of another. What is important in this connection is not the familiar $1/r^2$ - proportional static component of the gravitational force, but the acceleration-proportional radiative component of the interaction (Table 1). Einstein discussed this point a little in his book[2] in connection with the idealized experiment of Thirring. This description of the inertia of a given particle as arising from the radiative component of its interaction with all other masses in the universe has been looked into a little further by Sciama[3] and Davidson[4]. The inertial term ma is dropped from Newton's equation of motion. In its place appears the sum of the radiative interactions

$$ma \quad \sum_k \frac{G\, m_k\, f}{c^2\, r} \tag{1}$$

(2) A. Einstein, The Meaning of Relativity, Princeton University Press, Princeton, New Jersey, 3rd ed., 1950, p. 107; Scientific American, p. 209, April 1950.

(3) D. W. Sciama, Monthly Notices Roy. Astron. Soc. 113, 34 (1953); Scientific American, p. 99, February 1957.

(4) W. Davidson, Monthly Notices Roy. Astron. Soc. 117, 212 (1957)

This term gives a reasonable order of magnitude account of inertia if the dimensions of the universe are of the order of 10^{10} light years and if the effective average density of matter is of the order of 10^{-29} g/cm^3 (5).

Inertia is Tied to Geometry and Geometry is Directly Governed by the Distribution of Mass-Energy and Energy Flow

The analysis of Thirring and Einstein brings this "sum for inertia" into closer connection with the ideas of general relativity. On the one hand the inertial properties of a test particle are expressed in terms of the metric tensor $g_{\mu\nu}$. On the other hand the agencies responsible for changes in this measure of inertia are characterized not merely by density, but by the entire stress-energy tensor $T_{\mu\nu}$. Thus Thirring and Einstein write the change

$$h_{\mu\nu} = g_{\mu\nu} - \hat{g}_{\mu\nu}, \tag{2}$$

$$h = \hat{g}^{\mu\nu} h_{\mu\nu} \tag{3}$$

of the metric in a local Lorentz system, due to a change $\delta T_{\mu\nu}$, in the form

(5) For a discussion of present information on the density and size of the universe, see for example J. Oort and J. A. Wheeler in Onzième conseil de physique Solvay: La structure et l'évolution de l'univers, Editions Stoops, Bruxelles, 1959 (referred to hereafter as SÉU).

TABLE I. Static and radiative components of electromagnetic and gravitational forces compared and contrasted. The quantity f is an abbreviation for a dimensionless function of the angles between the lines of acceleration of source and receptor and the line connecting these two objects.

	Electromagnetism	Gravitation
Static or near part of interaction	$\frac{e_1 e_2}{r^2}$	$\frac{Gm_1 m_2}{f^2}$
Radiative or distant component	$\frac{e_1 e_2 a_2 f}{c^2 r}$	$\frac{Gm_1 m_2 a_2 f}{c^2 r}$

$$h_{\mu\nu} - \frac{1}{2} \hat{g}_{\mu\nu} h = (8\pi G/c^4) \int \frac{[\delta T_{\mu\nu}]_{ret} d^3x}{r} \qquad (4)$$

This expression remains a good approximate solution of Einstein's field equation so long as the geometry of the regions where the mass-energy is located does not differ substantially from the local Lorentz geometry at the position of the test particle. Looking at Eq. (4), and recalling that in relativity theory the inertial properties of a test particle are determined by the metric, one is led to formulate Mach's principle in the following form: (Formulation 2). <u>The geometry of space-time and therefore the inertial properties of every infinitesimal test particle are determined by the distribution of energy and energy flow throughout all space.</u>

<u>Many Objections to Mach's Principle</u>

That Mach's principle in anything like this form makes sense has been questioned on many sides for the following reasons:

(1) Einstein's field equations

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = (8\pi G/c^4) \; T_{\mu\nu} \qquad (5)$$

are non-linear. It is wrong in principle to try to express the solution $g_{\mu\nu}$ as a linear superposition of effects from the $T_{\mu\nu}$ in various regions of space.

(2) The quantity $1/r$ in the integrand is not a well defined quantity in an irregularly curved space.

(3) If in the Friedmann universe one considers the contributions to the inertia at a definite point in space-time from more and more remote points, where the retarded value of the stress-energy tensor is $T_{\mu\nu \mathrm{ret}}$, one is forced to go back to earlier and earlier moments of time. Ultimately one comes to a time when the system was in a singular state. What does one do then about the contribution of $T_{\mu\nu \mathrm{ret}}$ to the inertia!

(4) The elementary sum in (1) for the coefficient of inertia envisages a radiative interaction between particle and particle. But how can stars at distances of 10^9 and 10^{10} light years respond to the acceleration of a test particle here and now in such a way as to react back upon this test particle at this very moment? Is this difficulty not argument enough that this elementary formulation should be dropped? But when one turns from this picture of two-way travel of gravitational radiation to the Thirring-Einstein calculation where only one direction of travel comes into evidence, does one not encounter an ambiguity in this sense, that one could use advanced interactions just as well as

retarded interactions - or any combination of the two - in obtaining a solution of the linearized field equations? If the advanced and retarded expressions for the metric in terms of the distribution of mass-energy differ from each other - as expected - will not one be forced to conclude that one expression is wrong? And if one is wrong will it not be likely that both are wrong?

(5) Will not the 1/r - dependence of the supposed inertial interaction make the inertial properties of a test particle depend upon the expansion and recontraction of the universe, and the proximity of nearby masses, in a physically unreasonable way?

(6) How can it make sense to speak of the distribution of mass-energy (and energy flow) as determining the geometry? One cannot specify where one mass is, let alone the entire distribution of mass, until one has been given the geometry! But then what is there to be determined?

(7) Why spoil the beautiful logical structure of relativity theory by mixing up with it anything so vague and so lacking in mathematical sharpness as Mach's principle? Why try to word it in careful 20th century language when it is an outworn 19th century idea that ought to be dropped at once and for all time?

Solutions of Einstein's Equations not Produced but Selected by Mach's Principle

The answer is that Einstein's equations are not enough. Differential equations in and by themselves typically do not suffice to define a solution. They must be supplemented by a boundary condition. Mach's principle is required (Formulation 3) as a boundary condition to select allowable solutions of Einstein's equation from physically inadmissible solutions.*

---

*Note added after completion of this manuscript: This concept of Mach's principle as principle for the selection of solutions of Einstein's equations appears earlier in the discussion of J. A. Wheeler on pp. 49 - 51 of SEU, and especially in a recent article by H. Hön in E. Brüche, ed., Physikortagung Wien, Physik Verlag, Mosbach/Baden, 1962, where on p. 95 Hönl proposes two theses: (1) das Machsche Prinzip est als kosmologisches Prinzip ein Auswahlprinzip; d. h. es möglicher Lösungen des kosmologischen Problems einige wenigen aus zusondern, die als physikalisch sinnvolle Weltmodelle überhaupt in Frage kommen. (2) Das Machsche Prinzip lässt sich nur für räumlich geschlossene, endliche Weltmodelle in widerspruchloser Weise durchführen; es ist daher zu vermuten, dass die Forderung des Mach-Prinzips mit der Forderung eines endlichen Universums überhaupt identisch ist.

This kind of selection principle is so familiar in electrostatics (Table II) that it generally goes without even a name. Only when Poisson's equation is supplemented by such a boundary condition does it lead to the (1/r) law of action of a charge. This (1/r) law of action furnishes the usual basis for saying that the distribution of electric charge uniquely determines the distribution of electric potential.

<u>Cases Where the Boundary Condition Cannot be Applied Regarded as Idealizations of Cases Where It Does Apply and Where It Does Make Sense</u>

The boundary condition that the electrostatic potential shall fall off at large distances is noteworthy for what it does not do as well as for what it does do. It does not provide a way to <u>calculate</u> the (1/r)-law of action. Only the differential equation does that - giving in addition many other solutions. Moreover, one often considers in electrostatics problems where the requirement of Table II, "The potential must fall off at great distances" cannot be satisfied. By way of illustration, consider the problem: "Given $\rho(x,y,z) = \rho_0 \cos kz$; find $V(x,y,z)$" ! Thus one can choose between accepting the problem and giving up the generality of the boundary condition; or upholding the boundary condition at all times and modifying the problem.

TABLE II. Boundary conditions in electrostatic and in gravitation theory according to Formulation 3 of Mach's principle: a boundary condition to select allowable solutions of Einstein's equations from physically inadmissible solutions.

	Electrostatics	Gravitation Theory
Differential equations	$\nabla^2 V = -4\pi\rho$	The four of Einstein's equations which have to do with geometry on a space-like hypersurface.
Source terms	Electric charge density	Density of energy and energy flow.
General solution	$V = \int \frac{\rho d^3x}{r} + \Sigma c_{nm} r^n Y_n^{(m)}(\theta, \varphi)$	Geometry which (a) extends to spatial infinity or (b) is somewhere singular or (c) is closed up and free of singularity.
Principle of selection of physical solution	Potential must fall off at great distances	Geometry must be of class (c) (To admit singularities is to admit points where the equations are not really satisfied.
Consequence of this principle and also another way of formulating this principle.	Potential is uniquely determined by the distribution of charge	Geometry of <u>spacetime</u> must be uniquely determined by the distribution of energy and energy flow over the original <u>space-like</u> hypersurface.

One can say that the infinite cosine wave distribution of charge is only a mathematical idealization of a physical distribution of charge which is nearly of cosine character over a great region, as illustrated, for example, by an expression of the form

$$\rho(x,y,z) = \rho_0 \cos kz \exp[-(x^2 + y^2 + z^2)/a^2], \quad (6)$$

where the Gaussian breadth <u>a</u> is very large. On this choice of interpretation the boundary condition continues to make sense, and the potential continues to be determined uniquely by the distribution of charge.

<u>Asymptotically Flat Geometry Expressed as Limit of Closed Space</u>

Similarly in general relativity one can find situations which are not compatible with the boundary condition of Table II -- and therefore not compatible with formulation 3 of Mach's principle -- and which nevertheless can be translated over into situations which <u>are</u> compatible with the boundary condition. Consider for example a single spherically symmetric concentration of mass in otherwise empty space. Associated with this mass is the familiar Schwarzschild 4-geometry. This geometry is asymptotically flat at infinity. In this spacetime the inertial properties of an infinitesimal test particle approach indefinitely closely to the Newtonian expectations at indefinitely great distances from the mass. Consequently it is unreasonable to think of the central mass as responsible for these inertial properties. If one accepts this situation, he cannot uphold Mach's principle either as Mach originally stated it or as it is reformulated here, as a boundary condition to <u>select</u> solutions of Einstein's field equations:

(1) the inertial properties of test particles -- not being attibutable to the one mass that is present -- are therefore not assignable to Mach's "distribution of mass throughout all space;" and

(2) the Schwarzschild geometry does not describe a closed universe.

Therefore <u>rule</u> <u>out</u> around a center of mass a space that becomes flat at infinity. In other words, apply the geometric boundary condition of Table II to <u>exclude</u> the Schwarzschild geometry. Follow the example of electrostatics, where for example in Table II an infinite cosine distribution of charge was ruled out because it was incompatible with the boundary condition for the electrostatic potential at infinity.

The idealized situation that is pushed out of the back door as physically unacceptable comes in again at the front door in new clothes both in electrostatics and in general relativity. (Table III). Consider a geometry which is compatible with the boundary condition -- which is closed and free of singularity at some initial time, or more precisely on some initial spacelike hypersurface. To construct such a geometry, take not a single spherically symmetrical distribution of mass, but many such mass centers. Let the number of centers and their spacing be so chosen as to curve up the space into closure[7]. The dynamics of such a lattice universe before and after the

---

(7) For a detailed but approximate treatment of the dynamics of such a lattice universe, see R. W. Lindquist and J. A. Wheeler, Rev. Mod. Phys. 29, 432 (1957). For a precise analysis, consider the initial value problem at the moment of time symmetry or maximum expansion: ${}^{(3)}R = (16\pi G/c^2)\rho$. Here ρ is the density of mass, equal for example to ρ_o inside each center of attraction, and vanishing elsewhere. Solve this equation by modifying the geometry of a 3-sphere of uniform curvature and radius a,

$$ds^2_{ideal} = a^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\varphi)]$$

by a conformal factor Ψ:

$$ds^2 = \Psi^4 ds^2_{ideal}$$

The initial value equation takes the form

$$\nabla^2\Psi + (2\pi G/c^2)\rho\Psi^5 - (3/4a^2)\Psi = 0$$

Here the operator ∇^2 is calculated from the metric of the ideal moment of time symmetry agrees within a few percent or less with the dynamics of the Friedmann universe, with its filling by a uniform dust (zero pressure!) and its ideal uniform curvature. The corresponding expansion and recontraction of the lattice universe shows up, not so much through any change in the geometry interior to the typical Schwarzschild zone, as through a change in the place of join between one zone and the next. The interface moves outward from the centers of attraction on each side of it following the law of motion of a stone thrown out radially. It reaches a maximum distance. Then it falls back again towards both mass concentrations simultaneously. In this way the motion of these centers towards each other comes into evidence.

3-sphere. This equation is to be solved throughout one lattice zone subject to the conditions (1) that Ψ have the appropriate symmetry within that zone and (2) that its normal derivative vanish at the zonal boundary. This is an eigenvalue problem which determines the radius a of the comparison sphere. When gravitational radiation is present the metric cannot be represented in such a simple form. However, there is still typically a factor like Ψ to be found -- governed now not only by the distribution of mass, but also by the distribution of gravitational radiation.

Table III. Schwarzschild geometry envisaged as the limit of the geometry of a closed lattice universe when the size of the typical lattice zone is allowed to go to infinity. This limiting process is compared in the table with the analogous limiting procedure in electrostatics. Notation: (1) $m^*(\text{cm}) = (G/c^2)m(\text{g})$, mass at center of each lattice cell (2) $4\pi b^3/3$, volume of lattice cell at "instant" (spacelike hypersurface) of maximum expansion (3) a, radius of curvature of a comparison universe of uniform density and uniform curvature, also at the instant of mirror symmetry between past and future. This radius is determined as follows in terms of m* and b: The "Schwarzschild cells" are joined together on boundaries which are not sufficiently far out for the geometry there to be flat. The curvature of the Schwarzschild geometry in a local Lorentz frame in a plane perpendicular to the zonal radius is $R_{2323} = 2m^*/b^3$. Identify this quantity with the curvature in a typical plane in the uniform comparison universe, $R_{2323} = 1/a^2$. Thus find $a^2 \simeq b^3/2m^*$. Alternatively, write down the oo component of Einstein's field equations (the principal initial value equation of Yvonne Foures-Bruhat) in the form

$$^{(3)}R + (\mathrm{Tr}\ \underset{\sim}{K})^2 - \mathrm{Tr}\ \underset{\approx}{K}^2 = 2(8\pi G/c^4)\begin{pmatrix}\text{energy}\\ \text{density}\end{pmatrix}$$

Note that the extrinsic curvature tensor K_{ij} or "second fundamental form" vanishes on a time-symmetric spacelike hypersurface. Note also that the scalar curvature invariant of a 3-sphere of radius a, expressed in terms of the physical components (carat symbol!) of the curvature is:

$$^{(3)}R = {}^{(3)}\hat{R}_{11} + {}^{(3)}\hat{R}_{22} + {}^{(3)}\hat{R}_{33} = (\hat{R}_{1212} + \hat{R}_{1313}) + (\hat{R}_{2121} + \hat{R}_{2323}) + (\hat{R}_{3131} + \hat{R}_{3232}) = 6/a^2.$$

Identify the density of mass with $m/(4\pi b^3/3)$. Thus have $(6/a^2) \simeq (16\pi G/c^2)(3m/4\pi b^3)$ or again the result $a^2 \simeq b^3/2m^*$. The number of lattice cells is approximately

$N \simeq$ (volume of comparison universe)/(volume of cell)

$\simeq 2\pi^2 a^3/(4\pi b^3/3) = (3\pi/2)^{5/2}(b/m^*)^{3/2}$ (goes to infinity as size of typical cell goes to infinity).

	Electrostatic example	Example from general relativity
Source (before modification)	Infinite periodic charge distribution $\rho = \rho_0 \cos kz$	Single spherically symmetric concentration of mass in otherwise empty space
Effect of interest	Electric potential - and thence the electric field	Metric of spacetime - and thence the inertial properties of every infinitesimal test particle
Is "effect" so uniquely associated with "source" in this idealized case that one can say effect is "produced" by source?	No - can add to V any number of harmonics of form $r^n Y_n^{(m)}(\theta,\varphi)$	No - the asymptotically flat Schwarzschild geometry and many other empty space geometries solve Einstein's equations for this source "distribution".

	Electrostatic example	Example from general relativity
Does "effect" satisfy the boundary condition listed in Table II	No - none of these expressions for V falls off as fast as (1/r) at great distances	Schwarzschild geometry as normally conceived does not describe a closed universe.
Modified situation which <u>is</u> compatible with the boundary condition	$\rho = \rho_o \cos kz$ times $\exp(-r^2/a^2)$	Many such masses spaced with reasonable uniformity through a closed universe.
Scale factor associated with this new source	Range a of charge distribution	Effective radius b of typical Schwarzschild zone.
Is source now well defined?	Yes.	No. Must specify what gravitational waves if any are present-in other words, must specify otherwise undetermined features intrinsic to three geometry in which the masses are imbedded at the moment of time symmetry.*
When specification of "source" has been completed, is it reasonable to think of "effect" as well determined by this specification plus boundary condition?	Yes - in this event one can prove potential is uniquely determined by distribution of electricity.	Yes - expect other features intrinsic to this three geometry are now uniquely determined by (oo) component of Einstein's equation plus boundary condition of closure;**Mach's principle satisfied.

*See for example the "modified Taub universe" discussed in the text as an alternative to the lattice universe as a solution of Einstein's field equations which also satisfies the condition of closure.

**This uniqueness can be established in the case where the lattice universe contains no gravitational waves along the lines outlined in footnote 7. No investigation has been made of uniqueness when gravitational waves are present in this universe. However, there is a

	Electrostatic example	Example from general relativity
Limiting procedure now envisaged	Range a of charge distribution goes to ∞	Effective radius b of Schwarzschild zone goes to ∞
For each finite value of the parameter a or b is the relevant boundary condition satisfied?	Yes - V falls off as 1/r or faster at large r	Yes - Schwarzschild zone is a piece of a closed universe in which Mach's principle can be considered to apply.
Is boundary condition satisfied for infinite value of this parameter?	No - V does not fall off	No - Schwarzschild geometry is asymptotically flat.

related problem where the uniqueness of the 3-geometry - for specified distribution of gravitational radiation - has been established as a consequence of the closure condition[6].

(6) D. Brill, Ann. of Phys. 7, 466 (1959); H. Araki Ann. of Phys. 7, 456 (1959); J. A. Wheeler, Geometrodynamics, Academic Press, New York, 1962, p. 56. This book is cited hereafter as GMD.

The time for the expansion and contraction of the lattice universe - and of the boundaries of each Schwarzschild zone - is

$$\begin{pmatrix}\text{time for expansion}\\ \text{and recontraction}\\ \text{in length units}\end{pmatrix} = \pi \begin{pmatrix}\text{radius of lattice}\\ \text{universe at}\\ \text{maximum expansion}\end{pmatrix}$$

$$\simeq \pi \begin{pmatrix}\text{radius of one Schwarz-}\\ \text{schild zone at maximum}\\ \text{expansion}\end{pmatrix}^{3/2} \begin{pmatrix}\text{twice mass at center}\\ \text{of zone expressed in}\\ \text{length units}\end{pmatrix}^{-1/2} \qquad (7)$$

This quantity can be made arbitarily large relative to the time required for light to cross one Schwarzschild zone by making the radius b of the typical zone sufficiently large.

<u>Non-Uniform Convergence to Flat Space Limit</u>

The order of the participants is important. Let one participant, A, select (1) any arbitrary but finite distance from one center of mass and (2) any arbitrary but finite length of time and (3) any arbitrarily small but non-zero departure from the ideal Schwarzschild geometry which he is willing to tolerate. Then the other participant, B, can pick an effective radius for the typical Schwarzschild zone at the moment of maximum expansion which is so great that the geometry inside that zone agrees with the ideal Schwarzschild geometry (1) to within the specified limits of accuracy (2) out to the stated distance and (3) for the stated time. However, if B acts first, and specifies the zone radius at the moment of maximum expansion, then A can always point to places so far away that the geometry there totally disagrees with the continuation of the Schwarzschild geometry of the original zone. A can even point out that the space is closed and compatible with Mach's principle. Thus A concludes that the geometry is asymptotically flat or closed according as he is forced to make the first move or allowed to wait until B has fixed on dimensions. That A's conclusions depend upon the order of his move can be said in another way: The convergence to the limit of an infinitely great lattice universe is non-uniform.

Other Examples

The ideal lattice universe is no more than one of many conceivable examples to illustrate how one can consider as closed - and compatible in general terms with Mach's principle - geometries which ostensibly are asymptotically flat. Three more examples may give a slight impression of how wide is the range of allowable geometries.

Lattice Universe with Gravitational Radiation

In the lattice universe there may be present in addition to the "real" masses also the effective mass indirectly contributed by gravitational radiation. Then the inertial properties of test particles are affected by both sources of mass energy(7).

Modified Taub Universe

It is not necessary to supply any "real" masses additional to the one original mass in order to secure a closed universe. Gravitational waves of sufficient strength will supply the required curvature. This one sees from the example of the Taub universe(8). There gravitational radiation alone suffices to curve up the space into closure. In this 4-geometry consider the hypersurface or 3-geometry defined by the instant of time symmetry or maximum expansion. Perturb this geometry to the extent necessary to introduce a spherical ball of matter, at first arbitrarily small, eventually large or denser or both. Close to this mass the geometry is nearly Schwarzschildean. However, deviations from that limiting geometry become very great at distances comparable to the effective radius of the Taub universe(9). In this universe it is not reasonable to speak of a geometry primarily determined by "real mass" and perturbed in only a

---

(8) A. H. Taub, Ann. of Math. 53, 472 (1951)

(9) A first order analysis of deviations from Schwarzschild geometry has been given by T. Regge and J. A. Wheeler, Phys. Rev. 108, 1063 (1957), but no attempt is made there to fit on to the Taub solution a greater distances.

minor way by gravitational radiation. On the contrary, the gravitational radiation is the primary determiner of the 4-geometry - and on the inertial properties of test particles. The one "real mass" produces only minor perturbations in the geometry except in its own immediate neighborhood.

<u>Unmodified Taub Universe</u>

The fourth example is the Taub universe itself, free of any "real matter" at all. This solution of Einstein's equations for a closed empty space is interpreted in the appendix as a special case of a Tolman radiation filled universe in which (1) Tolman's electromagnetic radiation is replaced by gravitational radiation; (2) this gravitational radiation, instead of being effectively isotropic, is described by a single hyperspherical harmonic; and (3) this harmonic has the lowest possible order, or greatest possible wave length, compatible with the dimensions of the model universe.

<u>Does a Relation Between Inertia at One Place and Gravitational Radiation at Other Places Signify Circular Reasoning?</u>

Regardless of the details of the Taub universe, here is a closed space in which the inertial properties of every infinitesimal test particle are well determined. Yet there are no ordinary masses about, to interactions with which one can attribute the inertia of this test particle. Therefore, if Mach's principle is still to make sense, it is necessary to conclude that the distribution, not only of mass energy, but also of gravitational radiation, has to be specified in order completely to determine inertia - or, in the words of general relativity, completely to determine the geometry of spacetime. But gravitational radiation itself is described as an aspect of geometry and nothing more. Consequently one seems to be caught in a logical circle in trying to formulate Mach's principle. Apparently one has to give <u>the geometry</u> in advance, not only in order

(1) to say in any well defined way what one means by the term "distribution of mass-energy", but also

(2) to specify what gravitational radiation is present, so that one shall thereby be enabled

(3) <u>to determine the geometry</u> of spacetime!

Evidently one can never feel happy about a formulation of Mach's principle that seems to contain this kind of circular reasoning. Therefore it is essential to demand a mathematically well defined statement of his principle if Mach's ideas are to be considered as having any relevance at all for present day relativity physics.

Not Circular: Specify 3-Geometry, Determine 4-Geometry

Now for this mathematical formulation! It will be found to resolve the question of circular reasoning in this way, that what is specified is 3-dimensional geometry, and what is thereby determined is 4-dimensional geometry. At the same time it will help to clarify which features of gravitational radiation are freely disposable (field "coordinate" and its rate of change), and which features of the geometry are thereby determined (field "momentum").

II. 3-GEOMETRY AND ITS RATE OF CHANGE AS KEYS TO THE PLAN OF GENERAL RELATIVITY.

What is the "Plan" of General Relativity?

It is known often to help in answering one question to ask another. Therefore it is fortunate for the search for a mathematical formulation of Mach's principle - a search now physically motivated - that another issue is currently under discussion. As Professor J. L. Synge stated it at the Warsaw conference, what is the plan of general relativity? What quantities can one freely specify, and what quantities are thereby determined? What is the inner structure of the dynamic theory of a geometry governed by Einstein's field equations?

Plan 1: Initial Data on a Light-Like Hypersurface

One plan of dynamics starts with a light-like hypersurface.

In this approach as applied to the mechanics of a system of particles, one specifies the appropriate number of coordinates and momenta at the times when the respective world lines cross this null hypersurface. This formulation of mechanics has been investigated by P. A. M. Dirac and V. Fock. The corresponding formulation of geometrodynamics, particularly as relevant to the study of gravitational radiation, has been explored by R. Penrose, H. Bondi, R. Sachs and others, and has been described in a comprehensive report by Sachs at the Warsaw conference. However, this approach is not closely connected with the formulations of dynamics which are most widely used in other branches of physics. Whatever its relations with Mach's principle, they cannot be reported here because they have not been investigated.

<u>Plan 2: Coordinates and Momenta -- or Intrinsic Geometry and Extrinsic Curvature -- on a Space-Like Hypersurface</u>

Another plan of dynamics is more familiar. In particle dynamics give coordinates and momenta at points on the respective world lines which have a space-like relation each to the other. In electrodynamics give the field "coordinates" and "momentum" -- the magnetic field $\mathbf{B}(x^1, x^2, x^3)$ and the electric field $\underset{\sim}{E}(x^1, x^2, x^3)$ --everywhere on a space-like hypersurface. In geometrodynamics again give on a space-like hypersurface the field coordinates and momenta -- this time the 3-dimensional geometry intrinsic to this hypersurface,

$$ds^2 = {}^{(3)}g_{ik}(x^1,x^2,x^3)dx^i dx^k \quad , \tag{8}$$

and the "extrinsic curvature" or so-called "second fundamental form"[10] telling how this hypersurface is curved -- or to be curved --with respect to the enveloping -- or yet to be constructed -- 4-dimensional geometry. When the 4-geometry written in the form

$$d\sigma^2 = -d\tau^2 = {}^{(4)}g_{\alpha\beta}dx^\alpha dx^\beta = {}^{(3)}g_{ik}(x^o,x^1,x^2,x^3)dx^i dx^k + 2N_i dx^i dx^o + ({}^{(3)}g^{ik}N_i N_k - N_o^2)(dx^o)^2 \tag{9}$$

with the condition

$$x^o = x^{o*} \tag{10}$$

specifying the hypersurface in question, then the extrinsic curvature tensor is given by the expression[11]

---

(10) See for example L. P. Eisenhart, Riemannian Geometry, Princeton University Press, Princeton, New Jersey, 1926

(11) See R. Arnowitt, S. Deser and C. W. Misner, Phys. Rev. 122, 997 (1961) and earlier papers cited by them. This group of papers is referred to hereafter as ADaM. See also their chapter in L. Witten, editor, Gravitation: an introduction to current research, John Wiley, New York, publication scheduled for 1962. This book is referred to hereafter as GICR. See also P. A. Dirac, Proc. Roy. Soc. London A246, 333 (1958); Phys. Rev. 114, 924 (1959); Phys. Rev. Letters 2, 368 (1959).

$$K_{ik} = -\,(1/2N_o)(\partial\,{}^{(3)}g_{ik}/\partial x^o - N_{i|k} - N_{k|i}), \qquad (11)$$

in which x^o is understood as being fixed at the value x^{o*}. Here the vertical stroke is used to denote covariant differentiation with respect to the 3-geometry of the hypersurface, in contradistinction to the semicolon that marks covariant differentiation with respect to the 4-geometry. In terms of the extrinsic curvature tensor and its trace, the geometrodynamical momentum is[(12)]

$$\pi^{ik} = -\,({}^{(3)}g)^{\frac{1}{2}}(K^{ik} - {}^{(3)}g^{ik}\,\mathrm{Tr}\,\underset{\sim}{K}) \qquad (12)$$

<u>Interpretation of the Four Potentials or Metric Coefficients N_o and N_k as "Lapse Function" and "Shift Function"</u>

Some interpretation of the ADaM potentials N_α is appropriate. Imagine two thin ribbons of steel, distinguished from each other by the fact that one has painted on it the label $x^{o'} = 17.23$; the other, $x^{o''} = 17.27$. It is desired to construct out of these ribbons a rigid curtain. Paint cross-lines on the one ribbon at intervals which may gradually increase or gradually decrease but which never change regularly or erratically. Label them x' = 16,17,18... Do the same on

---

(12) This expression comes from ADaM.

the second ribbon, taking care that the new pattern of cross-lines is not widely different from the old pattern. Weld perpendicular uprights or "lapses" to the first strip at $x' = 16, 17, 18 \ldots$. As soon as these uprights have been cut to the right lengths, joined perpendicularly to the right points on the upper strip, and welded fast, the structure -- with all the curves thus forced into it -- will be determinate and rigid. To the waiting craftsman the architect sends two functions, $N_o(x')$ and $N'(x')$, the "lapse function" and the "shift function". The worker tabulates both at $x' = 16, 17, 18 \ldots$. In two further columns he tabulates for the same values of x' the product of N_o and of N' by the number $(x^{o''} - x^{o'}) = 0.04$. The one column tells him to what heights to cut off the uprights which he has welded to the strip that is lying down. The other tells him how far one way or the other to <u>shift</u> upper ends before he welds them to the upper strip. At $x' = 18$ let the value of what might loosely be called $N'dx^o$ be 0.5. This implies that the corresponding upright is welded at its bottom to the cross line marked $x' = 18$. The upper strip is shifted 0.5 <u>coordinate</u> units to the right. Thus the "lapse" is welded to it at a cross line marked $x' = 17.5$. How the "shift" changes from place to place -- and how much the spacing between one coordinate mark and the next differs between the upper and lower steel sheets -- together determine how much curvature is built into the curtain. Along this line of reasoning, generalized to three dimensions, one sees at once the reason for the mathematical structure of Eq. (11).

Interpretation in Terms of the Length of the Normal and the Difference in Space Coordinates at Its Two Ends

To state the same interpretations of N_o and N_k in other words, return to expression (9) for the distance between a point (x^o, x^1, x^2, x^3) that lies on one hypersurface, x^o = constant, and another point $(x^o + dx^o, \ldots, x^3 + dx^3)$ on another hypersurface, $x^o + dx^o$. Here the dx's are thought of as small but finite quantities. Let dx^o be kept fixed (at the value $dx^o = x^{o''} - x^{o'} = 0.04$, for example!) but on the hypersurface so selected let one point, then another, be tried until the invariant separation between it and the fixed point on the lower surface is extremized. Vary $d\sigma^2$ with respect to dx^k and set the coefficient of δdx^k equal to zero:

$$2\,{}^{(3)}g_{ik}dx^i + 2\,N_i dx^o = 0 \tag{13}$$

Solve for dx^i and find

$$dx^i = -{}^{(3)}g^{ik}N_k dx^o = -\,N^i dx^o \tag{14}$$

The extremal value of the separation comes out -- reasonably enough -- to be time-like:

$$d\tau = N_o dx^o. \tag{15}$$

Thus the "lapse function" N_o represents the _proper_ time separation between two hypersurfaces -- measured normally -- per unit of difference in their time _coordinates_. The vectorial "shift function" N^i represents the _coordinates_ at the base of the normal diminished by the coordinates at the summit of the normal, this difference again being referred to a unit difference between the time _coordinates_ of the two hypersurfaces.

Lapse and Shift Functions Required in Addition to 3-Geometry to Define 4-Geometry

Evidently it is not enough to specify the geometries ${}^{(3)}g_{ik}$ _intrinsic_ to a one parameter family of hypersurfaces in order to have a well defined 4-geometry. One must in addition tell how these hypersurfaces are _related_ to each other. One must tell how far apart the surfaces are ("lapse function") and how they are displaced space-wise one with respect to another ("shift function").

Arbitrary Lapse and Shift Functions Plus Arbitrary 3-Geometry Determine Field "Momentum"; but Arbitrary

Field "Momentum" and Arbitrary 3-Geometry are Ordinarily Incompatible.

From the field "coordinate" ${}^{(3)}g_{ik}$ and its rate of change with respect to the parameter x^0, plus information about the "lapse" and "shift" functions of position one can determine the "extrinsic curvature" K_{ik} and the associated field "momentum" (Eq. (11)). However, the converse is not generally true. If the field "coordinate" ${}^{(3)}g_{ik}$ and the field "momentum" or the extrinsic curvature K_{ik} are both specified arbitrarily, they will ordinarily be incompatible. The independent specification of the field coordinate and the field momentum is the wrong way to define initial value conditions in general relativity.

The Initial Value Equations

The incompatibility of arbitrary intrinsic geometry of field "coordinate" ${}^{(3)}g_{ik}$ with arbitrary extrinsic curvature or field "momentum" π^{ij} follows from four of Einstein's ten equations. These initial value equations[13] have to do with

---

(13) K. Stellmacher, Math. Ann. 115, 136 (1937); A. Lichnerowicz, J. Math. Pure Appl. 23, 37 (1944); Helv. Phys. Acta Supp. 4, 176 (1956); Theories relativistes de la gravitation et de l'electromagnetisme, Masson, Paris, 1955; Y. Foures-Bruhat, Acta Math. 88, 141 (1952); J. Rational Mech. Anal. 5, 951 (1956); and the chapter by Y. Foures in GICR.

conditions on the space-like hypersurface:

$$^{(3)}R + (\mathrm{Tr}\ \underset{\sim}{K})^2 - \mathrm{Tr}\ \underset{\sim}{K}^2 = 2(8\pi G/c^4)\begin{pmatrix}\text{energy}\\ \text{density}\end{pmatrix} \qquad (16)$$

$$(K_i{}^k - \delta_i{}^k\ \mathrm{Tr}\ \underset{\sim}{K})\Big|_k = (8\pi G/c^4)\begin{pmatrix}\text{density of flow of}\\ \text{energy in i-direction}\end{pmatrix} \qquad (17)$$

These initial value equations pose in sharpened form the issue, what is the plan of general relativity: what quantities

(1) can be freely and independently specified, and yet

(2) suffice completely to specify the past and future of the four-geometry?

Plan 3: Specify Completely Independently the Field Coordinates on Two Hypersurfaces

This question leads in turn directly to the two-surface formulation of dynamics, where one specifies no momenta, only coordinates (or conversely) -- but coordinates on two hypersurfaces rather than one.[14] Moreover, the field coordinates on the one surface are specified quite independently of those on the other surface. The complete freedom that one has in this way of specifying the initial value data would seem to be what one wants when he asks for a workable statement of the plan of general relativity (Table IV).

(14) The following is based on a paper of R. F. Baierlein, D. H. Sharp and J. A. Wheeler, Phys. Rev. 126, 1864 (1962), which in turn is based on (1) the A.B. Senior Thesis of David Sharp, Princeton University, May 1960 (unpublished) and (2) an analysis by R. F. Baierlein which led to the variational principle of Eq. (31).

Table IV. The plans of electromagnetism and general relativity as expressed in terms of the two surface formulation of dynamics. The field "coordinates" are specified on two space-like hypersurfaces--most simply on two hypersurfaces which have an infinitesimal separation.

	Electromagnetism	Gravitation
The physically significant field quantities	Components of the electromagnetic field	Components of the Riemann curvature tensor
The coordinate-independent object which they define	A 2-form: a honeycomblike structure of tubes of force	The intrinsic structure of the 4-geometry in the neighborhood (corrections to the Euclidean pattern of distances between one point and another in a great table of local "airline" (geodesic) distances
The dynamic equations which tell how this object changes from place to place	Maxwell's 8 equations	Equations that refer directly to the curvature components
The potentials normally introduced to simplify the analysis of these equations	The 4 components of the electromagnetic potential, A_α.	The 10 components of the metric tensor $g_{\mu\nu}$
Notation used for these potentials when spacetime is sliced into spacelike hypersurfaces	The magnetic potential A with components A_k and the electrostatic or scalar potential $\varphi = -A_o$	6 components of 3-metric $^{(3)}g_{ik}$ intrinsic to a slice; the normal proper time separation N_o between two hypersurfaces per unit of difference in their time coordinates; and the differences N^i (or more conveniently $N_k = {}^{(3)}g_{ki}N^i$ between

	Electromagnetism	Gravitation
		space <u>coordinates</u> at the two ends of such a normal, again per unit of difference in the time coordinates of the two hypersurfaces.
The dynamical problem as formulated in variational language for a region of spacetime bounded by two spacelike hypersurfaces σ' and σ''	Give $\underset{\sim}{A}'$ on σ' and $\underset{\sim}{A}''$ on σ''; in between take any trial functions $\underset{\sim}{A}(x^0,x^1,x^2,x^3)$ and $\varphi(x^0,x^1,x^2,x^3)$ calculate action integral; then vary the four potentials until the action is extremized.	Give ${}^{(3)}g'_{ik}\,(x^1,x^2,x^3)$ (this <u>defines</u> σ') and arbitrarily call the value of x^0 on this surface some number $x^{0'}$; similarly, give ${}^{(3)}g''_{ik}$ and $x^{0''}$. In between choose any trial values for the 10 potentials, compute action, extremize with respect to choice of the potentials.
The simpler version of this variational problem relevant for the formulation of initial value problem and Mach's principle: the two hypersurfaces have an infinitesimal separation.	Give $A(x^1,x^2,x^3)$ and $\partial A/\partial t$; have a simpler action principle in which $\varphi(x^1,x^2,x^3)$ is the only function to be adjusted.	Give ${}^{(3)}g_{ik}(x^1,x^2,x^3)$ and $\partial\,{}^{(3)}g_{ik}/\partial t$; have a simpler action principle in which only the "lapse function" $N_0(x^1,x^2,x^3)$ and the "shift function" $N_k(x^1,x^2,x^3)$ are to be varied.
Variational problem well defined in an <u>open</u> space?	No.	No.
Pay-off from this extremization in a <u>closed</u> space	Value of φ on the space-like surface from which one can then calculate the electric field $\underset{\sim}{E}$ - the "momentum" conjugate to the already specified field "coordinate" $\underset{\sim}{B}$.	Values of N_0 and N_k from which one can calculate the "extrinsic curvature" K_{ik} of the thin sandwich or the "momentum" conjugate to the geometrodynamical "coordinate" or intrinsic geometry ${}^{(3)}g_{ik}$.

	Electromagnetism	Gravitation
What equation has automatically been solved by this extremization?	The initial value equation div $\underset{\sim}{E} = 4\pi\rho$ in which there appeared superficially to be 3 unknown functions of position.	The initial value equations $(K_i{}^k - \delta_i{}^k \mathrm{Tr}\ K)_{\mid k} = (16\pi G/c^4)\hat{T}_{\perp i}$ $^{(3)}R + (\mathrm{Tr}\ \underset{\sim}{K})^2 - \mathrm{Tr}\ \underset{\sim}{K}^2 = (16\pi G/c^4)\hat{T}_{\perp\perp}$ in which there appeared ostensibly to be 6 unknown functions of position.
Situation now in brief	Have <u>compatible</u> values for field coordinate and field momentum on initial space-like hypersurface.	Have <u>compatible</u> values for field coordinate and field momentum on initial space-like hypersurface.
Further pay-off	Now have just the right amount of consistent initial value data to predict the electromagnetic field everywhere in space and at all times.	Now have just the right amount of consistent initial value data to determine the geometry of <u>spacetime</u> in past, present and future - and hence the inertial properties of every infinitesimal test particle.
Recapitulation of what information was required for this prediction	(1) Maxwell's equations (2) Law of motion of charges (3) Specification of divergence-free magnetic field and its time derivative on a closed space-like hypersurface (4) Specification of positions and velocities of charges at points where their world lines cross this hypersurface	(1) Einstein's equations (2) Dynamic law for the fields or objects responsible for the stress-energy tensor on the right side of Einstein's equations. (3) Specification of closed space-like 3-geometry and its rate of change with respect to a parameter x^o - a parameter which otherwise has no direct physical meaning.

	Electromagnetism	Gravitation
		(4) Specification of initial value data for fields or objects responsible for $T_{\mu\nu}$.
Is relation between "effect" and "source" well defined? (Mach's principle)	"Effect"-electromagnetic field. Relation well defined only if "source" is understood to imply specification on space-like hypersurface of both (1) positions of charges <u>and</u> (2) magnetic field and its time rate of change	"Effect"-inertial properties of test particle =geometry of spacetime. Relation well defined only if "source" is understood to imply specification on space-like hypersurface of both (1) density and flow of mass-energy <u>and</u> (2) intrinsic 3-geometry and its rate of change with respect to some parameter x^0 - this latter reasonably enough because how otherwise would one have a geometry with respect to <u>specify</u> the distribution and flow of mass.

Meaning of Phrase, "Independently Specifiable Coordinates"?

It is necessary to state in what sense one is to understand the phrase, "are specified quite independently of those on the other surface." What one says on this point depends upon the question whether he is thinking in the context of classical physics or quantum physics.

"Two-Surface" Formulation of Harmonic Oscillator Problem

By way of illustration consider the simple harmonic oscillator. Give the coordinate x' at the time t' and the coordinate x" at the time t". In this way fix the end points of a trial history,

$$x(t) = x_H(t) \tag{18}$$

The classical history in the intervening time interval is to be selected in such a way as to extremize the action integral

$$\begin{aligned} I_H &= \int_{x',t'}^{x'',t''} L(x_H(t), dx_H(t)/dt, t)\,dt \\ &= (m/2)\int (\dot{x}_H{}^2 - \omega^2 x_H{}^2)\,dt \end{aligned} \tag{19}$$

The solution is well known -- a simple harmonic motion of circular frequency ω:

$$x_H(t) = x_{H_{\text{classical}}}(t) = \frac{x'\sin\omega(t''-t) + x''\sin\omega(t-t')}{\sin\omega(t''-t')} \tag{20}$$

Associated with this "classical history" is the action --

"Hamilton's principal function" -- given by the expression

$$I_{H_{\text{classical}}} = [\ \omega/2 \sin \omega(t''-t')][(x'^2+x''^2)\cos \omega(t''-t')-2x'x''] \tag{21}$$

The Quantum Propagator and its Relation to the Classical Action

In quantum mechanics one gives arbitrarily, not the coordinates at two times, but the state function or probability amplitude $\Psi(x',t')$ at one time, t', and asks for its value $\Psi(x'',t'')$ at some later time t''. The function at the new time can be found by solving the Schroedinger equation numerically or otherwise. The focus of attention shifts from this equation to its solution in Feynman's formulation of quantum mechanics.[15] A propagator gives the desired function in terms of the arbitrarily specified initial function:

$$\Psi(x'',t'') = \int_{-\infty}^{+\infty} \langle x'',t''|x',t'\rangle\ \Psi(x',t')dx' \tag{22}$$

Feynman writes this propagator as the sum of elementary propagation amplitudes,

---

(15) R. P. Feynman, The Principle of Least Action in Quantum Mechanics, Ph.D. thesis, Princeton University, 1942 (unpublished); Rev. Mod. Phys. 20, 367 (1948); Phys. Rev. 76, 769 (1949); see also Philippe Choquard thesis, École Polytechnique Féderale, Zurich, 1955; and F. J. Dyson, Advanced Quantum Mechanics, photolithoprinted notes, Cornell University, Ithaca, New York, 1954.

$$\langle x'',t''|x',t'\rangle = \mathcal{N} \sum_H \exp\ (i\ I_H/h)\ . \qquad (23)$$

Every conceivable history contributes with the same weight; only the phase differs from one history to another. Destructive interference automatically cuts down the effective contribution of the non-classical histories. The sum reduces in the case of the harmonic oscillator to an expression of the form

$$\langle x'',t''|x',t'\rangle = \mathcal{N}_1 \exp\ (i\ I_{H\,\text{classical}}/h) \qquad (24)$$

where in the exponent Hamilton's principal function has the value (21).

Normal Compatible Versus Exceptional Incompatible Specification of "Two Surface" Data in Classical Problem

In the classical problem a difficulty arises when the time interval (t" - t') is an integral multiple of a half period. After an even number of half periods the coordinate must return to its initial value; after an odd number, it must come to the negative of its initial value. (1) If x" does not agree with x' in the one case, or with -x' in the other case, the end point data have been inconsistently specified. (2) Even if they have been consistently give, the momentum with which the motion starts off at the one end point -- and with which it returns to the other end point -- is completely indeterminate. In both cases the variational problem is indeterminate.

No Problem of Incompatibility in Quantum Propagator

No such problem of compatibility of the "end point data" or "two surface data" arises in the quantum formulation. When the interval (t"-t') is a half period, the propagator reduces to one type of Dirac delta function,

$$\langle x'',t''|x',t'\rangle = -\,i\,\delta(x''+x') \;; \tag{25}$$

and to another type when the interval is a full period:

$$\langle x'',t''|x',t'\rangle = -\delta(x''-x'). \tag{26}$$

In other words, the quantum propagator remains well defined for all specifications of the two surface data, regardless of any specialities in the classical problem in one case or another.

The Quantum Problem Always at the Background of Classical Analysis

No one has found any way to escape the conclusion that geometrodynamics, like particle dynamics, has a quantum character. Therefore the quantum propagator, not the classical history, is the quantity that must be well defined. Consequently it will not be considered a source of concern that one can specify the 3-geometries ${}^{(3)}\mathcal{G}'$ and ${}^{(3)}\mathcal{G}''$ intrinsic to two hypersurfaces in such a way that the action functional for general relativity admits no extremum. Such cases are the geometrodynamical generalization of the special cases just encountered for the harmonic oscillator. Only on this understanding will it be justified to say that the 3-geometry on one hypersurface is specified quite independently of the 3-geometry on the other hypersurface.

Concentration on the Case of Two Nearby Hypersurfaces

Of greatest simplicity is the case that alone will be considered here in any detail, where the two hypersurfaces are "close together". Then the determination of the momentum from the values of the coordinate on the two surfaces is the most immediate. This step carries one halfway through the dynamic problem. Having consistent and singularity free initial value data for momentum and coordinate at the initial time, one is in a position to complete the solution -- to determine without any ambiguity the history of the system for at least a finite proper time into the past and future[16]. For this purpose one uses the standard dynamical equations:

1. Hamilton's equations for a system of particles,
2. Maxwell's equations in the electromagnetic case,
3. Einstein's equations in the case of interest here.

Alternate Ways to Apply the Two-Surface Formulation of Dynamics

(16) The proof that this can be done in the case of general relativity is given in the book of A. Lichnerowicz, Théories relativistes de la gravitation et de l'électromagnétisme, Masson, Paris, 1955

Alternative ways of applying the two-surface formulation to particle mechanics, electrodynamics and general relativity differ from one another by the apportionment of the analytic load between a variational principle and differential equations.

Excluded Option 1: Well Separated Hypersurfaces and Exclusive Reliance Upon the Variational Method.

One can avoid any use at all of differential equations in calculating the history of the system, whether a particle, the electromagnetic field, or geometry. Instead one can rely entirely on the idea of extremizing an action integral extended over the entire interval of time for which one wants to know the history. For the particle, one specifies x' at t' and x'' at t''. One regards as the function to be varied, either $x(t)$ alone, as in the familiar Lagrangian variational principle of Eq. (19), or both $x(t)$ and $p(t)$ independently, as in the Hamiltonian formulation

$$\int_{x',t'}^{x'',t''} [p(t)\dot{x}(t) - \mathcal{H}(p(t),x(t),t)]dt = \text{extremum} \qquad (27)$$

To express electrodynamics in variational language one calls on the familiar vector and scalar potentials $\underset{\sim}{A}$ and φ,

$$\underset{\sim}{B} = \text{curl } \underset{\sim}{A}, \qquad (28)$$

$$E = -\partial \underset{\sim}{A}/\partial t - \text{grad } \varphi \qquad (29)$$

so that half of Maxwell's equations are automatically satisfied. The other four follow from extremizing the integral

$$I = \int [(1/8\pi)(\underset{\sim}{E}^2-\underset{\sim}{B}^2)+(\underset{\sim}{j}\cdot\underset{\sim}{A} - \rho\varphi)](1/c)d^4x \qquad (30)$$

Specified in advance are

(1) the charge and current densities ρ and $\underset{\sim}{j}$ (both in charge units/(length unit)^3) throughout the 4-dimensional region bounded by the two hypersurfaces

(2) $\underset{\sim}{B}$ on each of the two surfaces: in such a way that div $\underset{\sim}{B}$ vanishes -- this specification being made by giving $\underset{\sim}{A}$ on each of the two surfaces (arbitrary gauge; no effect on the physics from the change $\underset{\sim}{A} \rightarrow \underset{\sim}{A}$ + grad λ)

Varied everywhere between the two surfaces to extremize I are φ (quite independently) and $\underset{\sim}{A}$ (subject only to the specification of $\underset{\sim}{A}'$ and $\underset{\sim}{A}''$ at t' and t", respectively).

Option 1 Continued: The Variational Principle for General Relativity

The appropriate action principle in general relativity(14) --when supplemented by source terms--reads

$$I_4 = \int_{x^{o'},\, {}^{(3)}g_{ij}'}^{x^{o''},\, {}^{(3)}g_{ij}''} \Big\{\pi^{ij}\partial\, {}^{(3)}g_{ij}/\partial x^0$$

$$+ N_o({}^{(3)}g)^{\frac{1}{2}}[{}^{(3)}R - {}^{(3)}g^{-1}(\mathrm{Tr}\ \pi^2 - \tfrac{1}{2}(\mathrm{Tr}\ \pi)^2)] + 2N_i\pi^{ij}{}_{|j}$$

$$- N_o({}^{(3)}g)^{\frac{1}{2}}L^{**}(g^{\cdot\cdot},\ A\ldots.)\Big\}d^4x \qquad (31)$$

This variational principle results from adding complete derivatives to the familiar Lagrange integrand of general relativity, $({}^{(4)}R + L)(-g)^{\frac{1}{2}}$, and translating the result into the terminology of ADaM. Here L** is $8\pi G/c^4$ times the invariant or scalar Lagrangian for whatever fields have energy and produce gravitational effects, expressed in terms of (1) the covariant components of that field (the field components $F_{\alpha\beta}$ in electromagnetism for example) and (2) the elements $g^{\cdot\cdot}$ of the matrix reciprocal to $g_{\alpha\beta}$:

$$g^{\mu\nu} = \left\| \begin{array}{cc} ({}^{(3)}g^{jk} - N^j N^k/N_o^2) & (N^k/N_o^2) \\ & \\ (N^j/N_o^2) & -(1/N_o^2) \end{array} \right\| \tag{32}$$

Here ${}^{(3)}g^{jk}$ is in turn the matrix reciprocal to ${}^{(3)}g_{jk}$ and

$$N^j \equiv {}^{(3)}g^{jk} N_k \quad . \tag{33}$$

In (31) there are 16 functions of space and time to be varied in the region between the two surfaces in such a way as to extremize the integral. Ten of these quantities -- reasonably enough -- are metric coefficients: the six ${}^{(3)}g_{ik}$, free except for having to reduce to the prescribed values at the two surfaces; and the lapse and shift functions N_o and N_i (not N^i !), which are entirely freely disposable. The remaining six quantities, the momentum components π^{ij}, are also adjustable without any conditions at all. In spirit this adjustment of the momenta is like that of the particle momentum p(t) in Eq. (27). At the start the function is free even to the extent that its terminal values are free. However, extremization forces in that case the condition

$$\dot{x}(t) = \partial H(p,x,t)/\partial p \qquad (34)$$

from which the momentum is completely determined in terms of the velocity. Similarly here[11] ("Palatini philosophy"). Vary (31) with respect to π^{ij}. Set the variation equal to zero for arbitrary $\delta\pi^{ij}$. Find thus six equations determining the six π^{ij} in terms of the N_α and ${}^{(3)}g_{ik}$ and their derivatives. These equations are equivalent to Eq. (12) for the momentum in terms of the extrinsic curvature plus the definition of Eq. (11) for this extrinsic curvature. If one were concerned with translating the variational principle (31) back into differential equations, instead of using it as a variational principle, he would: (1) Vary the lapse and shift functions. (2) Set the coefficients of the four δN_α equal to zero. (3) Find in this way the four initial value equations (16,17) that have to do primarily with geometry within the successive hypersurfaces, (4) Obtain the other six more "dynamic" components of Einstein's ten field equations by varying the six ${}^{(3)}g_{ik}$ and setting the coefficients of the six $\delta\,{}^{(3)}g_{ik}$ equal to zero. But in using (31) in its alternative function -- to replace all differential equations (in the spirit of Rayleigh and Ritz) -- one will (1) substitute into (31) the expressions for the six π^{ij} in terms of the six ${}^{(3)}g_{ij}$ and the four N_α and their derivatives, and (2) use numerical methods or ten analytical trial functions $({}^{(3)}g_{ij}, N_\alpha)$ containing adjustable parameters to extremize the action integral I. Unhappily the extremum, rather than being a minimum or a maximum, is often a saddle of higher order, as one can convince himself even in the simpler problem of a single particle bound in a harmonic oscillator potential. This kind of variational principle does not normally lend itself either (1) at the theoretical level to establishing existence proofs or (2) at the practical level to doing calculations.

Most Favored Option 2: Two Infinitesimally Separated Hypersurfaces: Use of Variational Principle to Solve 2-Surface Initial Value Problem within the Thin Sandwich, then Einstein Field Equations to Predict All the Rest of the 4-Geometry; Electrodynamics as an Example

Proofs of the existence of solutions are much more widely known in manifolds with positive definite metric[17] than in

[17] C.B. Morrey, Pacific J. Math. 2, 25-53 (1952)
John Danskin, Rivista Mat. Univ. Parma 3, 43-63 (1952)

manifolds with indefinite metric. Moreover the real problem to be treated is the initial value problem. Once it has been solved one knows from the work of Lichnerowicz[16] that the solution can be continued by way of Einstein's ten field equations. Therefore concentrate on the thin sandwich problem. The essential ideas are most easily seen in the case of electromagnetism. The magnetic potential has been specified on both surfaces ($\mathbf{A}'$ and $\mathbf{A}''$) but the separation between them has been allowed to go to zero. Therefore the terms in $\mathbf{B}^2$ and in $\mathbf{j}\cdot\mathbf{A}$ <u>are not adjustable</u> in this limit. The variational principle reduces to the form

$$J = \int [(E^2/8\pi) - \rho\varphi] d^3x = \text{extremum} \tag{35}$$

to be extremized with respect to the single unknown potential φ. The theory of this variational problem is well known. Out of the extremization -- conducted analytically or by the Rayleigh-Ritz method or in any other way -- comes a potential φ that satisfies the differential equation

$$\nabla^2\varphi = -4\pi\rho - (\partial/\partial t)\,\mathrm{div}\,\mathbf{A} \tag{36}$$

This potential generates an electric field

$$\mathbf{E} = -\partial\mathbf{A}/\partial t - \mathrm{grad}\,\varphi \tag{37}$$

that automatically satisfies the initial value equation

$$\mathrm{div}\,\mathbf{E} = 4\pi\rho \tag{38}$$

One now has at hand $\underset{\sim}{E}$ and $\underset{\sim}{B}$ which can serve as the consistent starting points for the dynamic analysis. For this purpose apply the other six equations of Maxwell and predict the entire past and future of the electromagnetic field.

Concept of Thin Sandwich in Geometrodynamics

Similarly in relativity one seeks to adjust four potentials, the lapse function N_o and the shift function N_i, so as to generate an extrinsic curvature tensor K_{ij}, according to Eq. (11), which will satisfy initial value Eqs. (16,17). This done, the initial value problem is solved. To formulate the appropriate "thin sandwich" variational principle, proceed here as in electrodynamics to the limit in which the sandwich is indefinitely thin. One can state this idea in two alternative ways[14]. (1) Give nearly identical ${}^{(3)}g'_{ik}$ and ${}^{(3)}g''_{ik}$. Take any arbitrary numbers $x^{o'}$ and $x^{o''}$ for the labels to be applied to these two hypersurfaces. In the definition of the extrinsic curvature K_{ik} (Eq. (11)) there enters the term $\partial\,{}^{(3)}g_{ik}/\partial x^o$. Adopt for this term the value $({}^{(3)}g_{ik}'' - {}^{(3)}g_{ik}')/(x^{o''} - x^{o'})$. Apparently the value of K_{ik} will depend on $(x^{o''} - x^{o'})$. Actually it will not. All that ever matters in K_{ik} or anywhere else is the <u>product</u> of $(x^{o''} - x^{o'})$ by the lapse function N_o. If a big value is used for $(x^{o''} - x^{o'})$, a small value will come out of the variational principle for N_o, and conversely. One sees this invariance property of the product also in another way, that the normally measured interval of <u>proper</u> time between the two hypersurfaces (Eq. 15) is $N_o(x^{o''} - x^{o'})$. Therefore in this formulation one takes as the quantities to be varied only the <u>products</u>

$$\eta_o = N_o(x^{o''} - x^{o'}) \qquad 39)$$

$$\eta_k = N_k(x^{o''} - x^{o'}) \qquad (40)$$

and never lets the individually arbitrary quantities N_o, N_k, $x^{o'}$, $x^{o''}$ show up. To this conceptually simpler formulation

hat is kept fixed during the variation (${}^{(3)}g_{ik}{}'$ and ${}^{(3)}g_{ik}{}''$) there is an alternative and mathematically sharper statement. (2) Consider a continuous one-parameter (x^o) family of 3-geometries ${}^{(3)}g_{ik}(x^0,x^1,x^2,x^3)$. Then the initial value problem under consideration is defined by a knowledge of ${}^{(3)}g_{ik}$ and $\partial\,{}^{(3)}g_{ik}/\partial x^o$ for some one fixed value of x^o. The associated variational problem is found by dropping the factor dx^o in the integrand d^4x in (31).

<u>The "Intrahypersurface Variational Principle" for the Initial Value Problem of General Relativity</u>

Now that only a three-fold space integration is called for, the next to the last term in (31) can be integrated by parts:

$$2N_i \, \pi^{ij}{}_{|j} \rightarrow - \pi^{ij}(N_{i|j} + N_{j|i}). \tag{41}$$

In a non-Euclidean topology more than one coordinate system is generally required to cover a manifold without singularity. Each is defined in its own coordinate patch.[18] It might appear that a problem of transition arises in passing from one patch to another in the integration by parts. The absence of any such difficulty is guaranteed by the covariant character of the differentiations in (41). Moreover, the surface integral disappears in the simplest example of a closed space, a manifold with the topology of the 3-sphere S^3. Thus, let the integration start in the neighborhood of one point P in S^3. Let it extend out to a boundary with the topology of the 2-sphere S^2. As the range of integration is widened, S^2 at first swells more and more. Later it begins to decrease in size. Eventually, as the integration extends over the whole 3 space, the boundary collapses to nothingness at some point other than P. No surface integral is left. Also no derivatives of the π^{ij} are left in (31).

(18) See for example GMD, p. 259

Therefore everywhere that these momenta appear, they are easily expressed in terms of the curvature tensor K_{ij} by (12); and the K_{ij} are then expressed -- via (11) -- in terms of the quantities that one really thinks of varying: the lapse and the shift functions. The first substitution leads by simple algebra to the formula,

$$I_3 = \int \left\{ {}^{(3)}R - (\mathrm{Tr}\ K)^2 + \mathrm{Tr}\ K^2 - L^*(g^{\cdot\cdot}, A \ldots) \right\} \cdot ({}^{(3)}g)^{\frac{1}{2}} N_0 d^3x \tag{42}$$

for the quantity to be extremized. In this "intrahypersurface" (IHS) variational principle, as in other applications of the Lagrangian method to dynamics, the "kinetic" term $(\mathrm{Tr}\ K)^2 - \mathrm{Tr}\ K^2$, appears with a sign opposite to that of the "potential" term ${}^{(3)}R$, whereas in the initial value equation (16) for the <u>energy density</u> these terms appear - reasonably enough - with the same sign. The second substitution - writing the K_{ij} and the $g^{\alpha\beta}$ in terms of the four functions to be varied, the N_α, by using Eqs. (11) and (32) -- is better left understood than carried out explicitly!

<u>The Also Useful Option 3: Exclusive Reliance on Differential Equations to Analyze the Dynamics of General Relativity</u>

Option 3 for analyzing the "plan" of general relativity, like option 2, starts with a specification of $^{(3)}g_{ik}$ and $\partial\,^{(3)}g_{ik}/\partial x^{o}$ over the entirety of a closed space-like hypersurface; in more picturesque language, it presumes a specification of two "nearby" 3-geometries $^{(3)}\mathcal{G}'$ and $^{(3)}\mathcal{G}''$. Here "nearby" is to be tested after the event by calculating N_o and from it (Eq. (15)) finding if the proper time separation between the two hypersurfaces is or is not small compared to the scale of the space-like variations in $^{(3)}\mathcal{G}'$ and $^{(3)}\mathcal{G}''$. In addition, the density of energy -- and energy flow -- have to be given, just as in Option 2. The difference is only that the four potentials N_α are to be found by solving the four Eqs. (16,17) -- not by directly trying to extremize the action integral I_3 of (42). Once the lapse and shift have been found, however, there is no difference in what one does between Option 3 and Option 2. (1) Calculate the extrinsic curvature K_{ik}. (2) Calculate the field momentum π^{ik}. (3) Use all ten of Einstein's equations to predict the 4-geometry in past and future.

<u>Verification that the Intrasurface Variational Principle and the Initial Value Equations are Equivalent</u>

On the right hand side of the initial value equations stand the density of energy and energy flow, a total of four quantities. In contrast, the variational principle (42) makes reference to all of the covariant components of the field responsible for this energy. One could therefore be concerned whether the two approaches will give the same result. To check this point, vary the N_α in the variational principle of (42). Set the coefficients of the δN_α equal to zero. Finally compare with the initial value equations. The variation of the field Lagrangian is the most complicated part of this program. Write

$$\delta[N_o L^*(g^{\cdot\cdot}, A\ldots)] = L\delta N_o + N_o(\partial L^*/\partial\,^{(4)}g^{\alpha\beta})(\partial\,^{(4)}g^{\alpha\beta}/\partial N_\gamma)\delta N_\gamma \tag{43}$$

Evaluate the derivatives of the components of the reciprocal metric tensor by using Eq. (32) for that tensor. Express the derivatives of the Lagrange function in terms of the stress-energy tensor of the field in question, employing for this purpose the standard formula,[19]

$$\begin{aligned} T^{**}_{\alpha\beta} &= (-g)^{-\frac{1}{2}}(\partial/\partial g^{\alpha\beta})(-g)^{\frac{1}{2}}\,L^{**} \\ &= (\partial L/\partial g^{\alpha\beta}) - (1/2)g_{\alpha\beta}L^{**} \end{aligned} \tag{44}$$

Here $T^{**}_{\alpha\beta}$ (m^{-2}) is an abbreviation for ($8\pi G/c^4$) times the usual stress-energy tensor $T_{\alpha\beta}$ (kg m^2/sec^2 m^3). Find that all those

---

(19) See for example L. Landau and E. Lifshitz, translated by M. Hammermesh, The Classical Theory of Fields, Addison-Wesley Press, previously Cambridge, now Reading, Massachusetts, 1951.

terms in (43) go out which contain an undifferentiated L factor. Those that remain give

$$2[T^{**}_{\underline{11}}\ \delta N_o + T^{**k}_{\underline{1}}\ \delta N_k] \qquad (45)$$

Here

$$T^{**}_{\underline{11}} \equiv (T^{**}_{oo} - 2N^k T^{**}_{ok} + N^i N^k T^{**}_{ik})/N_o^2$$

$$= T^{**\underline{11}} = (8\pi G/c^4) \left(\begin{array}{l}\text{density of energy as corrected} \\ \text{for the ordinarily oblique} \\ \text{coordinate system in use, a} \\ \text{scalar with respect to coordi-} \\ \text{nate changes in the hypersur-} \\ \text{face.}\end{array}\right) \qquad (46)$$

and

$$T^{**k}_{\underline{1}} \equiv {}^{(3)}g^{km}(T^{**}_{om} - N^s\ T^{**}_{sm})/N_o;$$

$$T^{**k\underline{1}} = T^{**k}_{\underline{1}} = (8\pi G/c^4) \left(\begin{array}{l}\text{density of flow of energy,} \\ \text{corrected for oblique co-} \\ \text{ordinate system off sur-} \\ \text{face, a contravariant vector} \\ \text{with respect to coordinate} \\ \text{changes in the hypersurface}\end{array}\right) \qquad (47)$$

The rest of the variational analysis is straighforward. One verifies the agreement with the initial value equations in all detail.

Precisely What Features of the Energy are Specified on the Hypersurface?

As the quantities which are specified on the hypersurface in the *initial value equations* one evidently thinks most naturally of

$T^{**}_{\perp\perp}$ and $T^{**}_{\perp}$, not the much more coordinate dependent $T^{**}_{\alpha\beta}$. As regards the variational principle, it is clear that it can be changed -- if only the change reproduces the initial value equations. Therefore the Lagrange function, which may be complicated or unknown or both, can be replaced by an expression which will have the same variation (45). Thus one comes to the modified variation principle,

$$I^*_3 = \int \{[{}^{(3)}R - (\mathrm{Tr}\underset{\sim}{K})^2 + \mathrm{Tr}\underset{\sim}{K}^2 - 2T^{**}_{\perp\perp}]N_o - 2T^{**k}_{\perp} N_k\}({}^{(3)}g)^{\frac{1}{2}}\, d^3x \tag{48}$$

Elimination of the Lapse Function

The lapse function N_o enters only algebraically in the time component (16) of the initial value equations and in the variational principle (48). To bring this fact most clearly into evidence, introduce the abbreviation

$$\gamma_{ij} = (1/2)[N_{i|j} + N_{j|i} - \partial\, {}^{(3)}g_{ij}/\partial x^o] \tag{49}$$

and write

$$\gamma_2 = (\mathrm{Tr}\ \underset{\sim}{\gamma})^2 - \mathrm{Tr}\ \underset{\sim}{\gamma}^2 \tag{50}$$

("shift anomaly"). Then

$$K_{ij} = \gamma_{ij}/N_o \tag{51}$$

K_{ij} measures the true extrinsic curvature, having to do with changes in space-like distances per unit of <u>proper</u> time between two hypersurfaces. In contrast, γ_{ij} performs a similar function when one does not yet know the lapse function, or scale of proper time, so that one has to use a purely nominal time coordinate x^o. The "kinetic" term in the variational principle becomes

$$(\mathrm{Tr}\,\underset{\sim}{K})^2 - \mathrm{Tr}\,\underset{\sim}{K}^2 = \gamma_2/N_o^{\,2} \tag{52}$$

The modified variational principle becomes

$$I_3^* = \int\{({}^{(3)}R - 2T_{\perp\perp}^{**})N_o - \gamma_2/N_o) - 2T_{\perp}^{**k}N_k\}({}^{(3)}g)^{\frac{1}{2}}d^3x \tag{53}$$

If there exists an extremum with respect to N_o, it occurs for

$$N_o = [\gamma_2/(2T^{**} - {}^{(3)}R]^{\frac{1}{2}} \tag{54}$$

The opposite sign for the root gives nothing physically new. With this reversal in sign N_k also comes out reversed in sign. All that has been changed is the convention as to the direction in which time is increasing! Reference (14) comments about the result (54): "Thus not only is the thickness of the thin sandwich from ${}^{(3)}\mathcal{G}'$ to ${}^{(3)}\mathcal{G}''$ determined by ${}^{(3)}\mathcal{G}'$ and ${}^{(3)}\mathcal{G}''$, but also its location in the enveloping ${}^{(4)}\mathcal{G}$ is determinate. This is the sense in which we discover a 3-geometry to be the carrier of information about time in general relativity."

The Condensed Intrasurface Variational Principle as Mathematical Formulation of Mach's Principle

Insert expression (54) for the lapse into (53) and obtain the "condensed intrasurface variational principle"[20] (CIVP),

$$\boxed{\begin{aligned} I_{CIVP} = -I_3^*/2 &= \int\{[\gamma_2(2T_{\perp\perp}^{**} - {}^{(3)}R)]^{\frac{1}{2}} + T_{\perp}^{**k}N_k\}({}^{3}g)^{\frac{1}{2}}d^3x \\ &= \text{extremum} \end{aligned}} \tag{55}$$

The analogue of this intrasurface variational principle in electrodynamics is

$$\int [(\underset{\sim}{E}^2/8\pi) - \rho\varphi] d^3x = \text{extremum} \tag{56}$$

equivalent with

$$\underset{\sim}{E} = - \partial A/\partial t - \text{grad}\ \varphi \tag{57}$$

to the single differential equation

$$\nabla^2\varphi = - 4\pi\rho - (\partial/\partial t)\text{div}\ \underset{\sim}{A} \tag{58}$$

for the single potential φ. In (55) the given quantities are still the metric ${}^{(3)}g_{ik}$ of the hypersurface, the rate of change of this metric with a parameter x^0, the scalar curvature invariant ${}^{(3)}R$ of the geometry, and the density of energy and energy flow. To be varied to obtain an extremum are now not four potentials but only three, the components N_k of the vectorial shift function. They enter (55), (1) as coefficients of the energy flow and (2) as determiners -- through their covariant derivatives of the "shift anomaly" γ_2. The variational principle CIVP of (55) expresses in precise mathematical form the principle of Mach as formulated here (Formulation 4): <u>the specification of a sufficiently regular closed 3-dimensional geometry at two immediately succeeding instants, and of the density and flow of mass-energy, is to determine the geometry of spacetime, past, present and future, and thereby the inertial properties of every infinitesimal test particle</u>. Thus from (55), when it possesses a solution, one obtains the shift N_k. Then from (54) one has immediately the lapse function. From these potentials via (49) and (51) one obtains the extrinsic curvature. Then one has in hand all the initial value data -- and consistent initial value data -- which one needs for the integration of Einstein's field equations and for obtaining a uniquely specified 4-geometry (the arbitrariness in the coordinate system in this spacetime having no relevance to its geometry!)

Condensed Initial Value Equations

Make small variations δN_k in the shift components in (55). Set the coefficients of these variations equal to zero. In this way arrive at three coupled second order differential equations for the determination of the vector field $N = (N_1, N_2, N_3)$. The same equations may be obtained by solving (16) for N_o (in agreement with (54)) and substituting this result into (17). The condensed initial value equations read

$$\left\{ \frac{(2T^{**\perp\perp} - {}^{(3)}R)^{\frac{1}{2}} [(N_{i|j} + N_{j|i} - {}^{(3)}\dot{g}_{ij}) - {}^{(3)}g_{ij}\, {}^{(3)}g^{mn} (N_{m|n} + N_{n|m} - {}^{(3)}\dot{g}_{mn})]}{[({}^{(3)}g^{ab}\, {}^{(3)}g^{cd} - {}^{(3)}g^{ac}\, {}^{(3)}g^{bd}) (N_{a|b} + N_{b|a} - {}^{(3)}\dot{g}_{ab}) (N_{c|d} + N_{d|c} - {}^{(3)}\dot{g}_{cd})]^{\frac{1}{2}}} \right\}_{|j}$$

$$= - T^{**}_{\perp i} = + (8\pi G/c^4) \left(\begin{array}{c} \text{i-th covariant component of} \\ \text{density of flow of energy} \end{array} \right) \qquad (59)$$

Variational Principle Equivalent to Differential Equations Plus Boundary Conditions

These equations plus boundary conditions are equivalent to the condensed intrasurface variational principle (55). The <u>boundary conditions are essential</u> in geometrodynamics as in electrostatics <u>if</u> one is to obtain a unique relation between the "source" (density and flow of energy and gravitational radiation as described by ${}^{(3)}g_{ij}$ and ${}^{(3)}\dot{g}_{ij} = (\partial\, {}^{(3)}g_{ij}/\partial x^{0})$ and the "effect" (the vector shift $\underset{\sim}{N}$ and the 4-geometry and inertial properties of test particles). The boundary conditions in a closed space are obvious: the vector field $\underset{\sim}{N}$ found by integration around the space one way has to join up properly with the vector field found by integration around the space another way; or more simply, the vector field (due account being taken of changes from one coordinate patch to another[18] (1) must be everywhere <u>regular</u> and (2) must lead to a regular and <u>single valued</u> extrinsic curvature K_{ij}. If the space is <u>open</u>, the differential equations are still well defined; but they are not accompanied by any boundary condition. Moreover, one can no longer expect the variational integral ordinarily to have a finite and well defined value in the case of an open space. Therefore there arises the built-in consequence of Mach's principle as formulated here, that the space should <u>be closed</u> and that the geometry (${}^{(3)}\mathcal{G}'$ and ${}^{(3)}\mathcal{G}''$, or ${}^{(3)}g_{ik}$ and $\partial\, {}^{(3)}g_{ik}/\partial x^{0}$) should be everywhere regular.

III. COMMENTS ON MACH'S PRINCIPLE AND THE INTRASURFACE VARIATIONAL PRINCIPLE

Issues Not Discussed Here: Uniqueness

It would be an enormous labor to take up one by one all the questions that are left unanswered here and treat them systematically. Moreover, there is wanting one key element in the discussion -- a proof that the solution of the variational problem in (55) (when there is a solution) is unique[20].

---

(20) The question of uniqueness of the solution of the initial value problem is well understood in the case of electrodynamics in a closed orientable 3-manifold. Given everywhere $\underset{\sim}{B}$ and $\dot{\underset{\sim}{B}}$, one only then arrives at a unique $\underset{\sim}{E}$ when one specifies the jump $\Delta_k\varphi$ in the potential in travelling the circuit of the k^{th} independent handle or "wormhole" of the topology, where k runs over the values from $k = 1$ to $k = R_1 = R_2$ = the second Betti number of the manifold. These numbers determine the charge or flux of lines of force trapped in the topology. That the numbers $\Delta_k\varphi$ have to be fixed follows most evidently from the occurence of a surface integral $\int\delta\varphi(\underset{\sim}{E}\cdot d\underset{\sim}{S})$ in the passage from the variational principle (35) to the differential equation (36). Does topology make an equally forceful appearance in the initial value equations of general relativity? <u>Is there a geometrodynamical analogue of</u>

<u>electric charge? No argument for the existence of such a charge follows from the variation principle as discussed in the text ("coordinate representation")</u> The surface integral of the quantity $\pi^{ij}N_i$ shows up in the integration by parts of Eq. (41). In the discussion of the text following that equation it is remarked that this surface integral vanishes when the topology is that of a 3-sphere (no handles). However, the surface integral <u>also</u> vanishes (C. W. Misner) for <u>any</u> closed orientable 3-manifold. The nature of the 2-surfaces encountered in these integrals is the same in geometrodynamics as in electromagnetism. Most simply, one such surface is conceived as the point of contact between two balloon-like expanding fingers that are feeling their way down into a wormhole from opposite mouths. The first factor in each integrand -- $\underset{\sim}{E}$ in the one case, π^{ij} in the other case -- is the same in this respect, that the quantity in question has <u>physical meaning</u> and is a <u>field momentum</u>. The difference comes in the character of the second factor -- the potential jump $\delta\varphi$ in electrodynamics, the metric potential N_i in geometrodynamics. Only the <u>gradient</u> of φ has significance in electromagnetism, so that φ itself can suffer a net change in going around the circuit of a handle. On the other

hand, the quantity N_i directly governs the distance between points on the two nearby hypersurfaces that have specified coordinates. Unlike the electric potential φ, this quantity must return to its original value after the circuit of a handle. Therefore a geometrodynamic analog to electric charge -- if one is to come in at all -- will have to show up in the conjugate representation of the initial value problem (not analyzed here).

<u>Effect of Additional Mass on Inertia not Discussed</u>

On the other side of the story there are many homely questions about the physical content of Mach's principle that ought to be spelled out and that now can be spelled out. An example is the question how the inertial properties of a sun and planet are affected if centered around them at some distance is constructed a very large spherical shell of mass. Here it is necessary to recognize that in one way the inertial properties are affected and in another way they are not, according as the clocks in use are within the shell or far outside it. Again, subtleties arise which are better left unmentioned than discussed inadequately.

<u>Instantaneous or Retarded Effect of Source on Test Particle</u>

Another question has to do with the <u>speed</u> with which the supposed inertial effects of sources are propagated to the test particles which they affect. In the equation (58) connecting source and effect even in electrodynamics, the effects of the charge distribution on the potential appear formally to be propagated instantaneously within the space-like hypersurface. Yet the whole analysis goes back to standard Maxwell electrodynamics, in which effects are all propagated, not instantaneously, but with the speed of light. That there is no inconsistency between the instantaneous potential of (58) and the retarded potentials of usual radiation theory is well known.[21] Analogously one finds also in geometrodynamics a basically elliptic equation, describing what appears formally to be an instantaneous propagation of effects from one place to another in a spacelike hypersurface. Yet one knows that a disturbance in a source at one point in spacetime will propagate to another point only with the speed of light.[22] In geometrodynamics as in electrodynamics the formalism itself guarantees that there can be no discrepancy between effects calculated in the two different ways from the same sources.

---

(21) E. Fermi, Rev. Mod. Phys. 4, 87 (1932).

(22) Marcel Riesz, Acta Mathematica 81, 1, 223 (1949)

Therefore in principle there can be no trouble from the e question mentioned earlier: How can Mach's principle make sense when it implies that the accelerated test mass acts on all the other masses in the universe, and that they in turn have to act back on this particle?[23] Of course one would like here, as in Fermi's analysis of electrodynamics, to see more of the inner workings of the machinery by which (1) the propagation in time and (2) a formally instantaneous propagation necessarily yield the same solution of Einstein's field equations!

Do Sources Have to be Followed Back into Past when Model Universe Was in a Singular State?

That all effects appear formally as propagated instantaneously within the space-like hypersurface disposes of another question about Mach's principle. Let one evaluate the inertial effects on a given test particle -- that is to say, the effects on the geometry in a given neighborhood -- caused in the sense of Mach by more and more remote sources of mass-energy. One appears to be forced farther and farther

(23) For more on the equivalence between retarded and other ways of evaluating potentials in electrodynamics, see for example J. A. Wheeler and R. P. Feynman, Rev. Mod. Phys. 17, 157 (1945) and 21, 425 (1949).

back in the past. On this basis one ultimately comes to regions where the geometry is singular and where it is not possible to follow back any further the dynamical evolution of the geometry by employing Einstein's field equations only at the classical level.(24) No matter! Specify the dynamic problem by giving the "sandwich" type of data on an initial space-like hypersurface: give ${}^{(3)}\mathcal{G}$, $\partial\,{}^{(3)}\mathcal{G}/\partial x^{o}$, and the density and flow of energy. Then the integral that one has to extremize or the triplet of differential equations that one has to solve make no reference to anything going on back in the past at a time or place where the geometry -- calculated classically -- may be singular.

<u>Model Universe Clean of Constants of Motion</u>?

Still another question is this, "what are the true physical constants of the motion" in general relativity. It is well known that total energy cannot be defined and has no meaning in a closed universe.(19) The question has recently been raised(24) whether such a system is not in principle <u>clean</u> of all constants of motion whatsoever. One can compare a model universe in some respects with a billiard ball set into

(24) This question of singularities is raised and discussed further in an article by the author in press in the special cosmology issue of The Monist, Vol. 47, No. 1, Box 268, Wilmette, Illinois.

motion on a triangular billiard table which has sides e, π and 1. The motion is quasiergodic. Started in one way the billiard ball will come indefinitely close to repeating the motion it would have had if it were started in another way. To an observer with only a finite resolving power the only difference in the two motions might be one of rate or energy. Not even this difference can manifest itself in the case of a model universe.(24) Nevertheless, there is no more difficulty in defining the dynamics of the billiard ball (by giving x', y' at t' and x", y" at t") than there is in defining the dynamics of geometry (by giving ${}^{(3)}\mathcal{G}'$ and ${}^{(3)}\mathcal{G}''$). In other words, if there are no constants of the motion they will hardly be missed!

Different Masses on the Two Hypersurfaces

Now for questions on which something more definite can be said. First, how can it possibly make sense to specify ${}^{(3)}\mathcal{G}'$ and ${}^{(3)}\mathcal{G}''$ arbitrarily? Are there not all sorts of conditions of compatibility that have to be satisfied? Consider for example the case of a space that is asymptotically flat. From the rate of approach to flatness at great distances,

$$ds^2 \sim (1 + 2m^*/r)dr^2 + r^2(d\theta^2 + \sin^2\theta \, d\varphi^2) \qquad (60)$$

one can evaluate the mass and energy of the system. If this has to be the same on both hypersurfaces, how many other constants must there not also be which have to agree between ${}^{(3)}\mathcal{S}'$ and ${}^{(3)}\mathcal{S}''$? To discuss this question more fully, consider a specific example, the Schwarzschild solution of Einstein's field equations,

$$d\sigma^2 = -d\tau^2 = -(1 - 2m^*/r)dt^2 + (1 - 2m^*/r)^{-1}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\varphi^2) \qquad (61)$$

Let ${}^{(3)}\mathcal{S}'$ be the hypersurface t = t' = constant. On this the asymptotic geometry follows Eq. (57). Let the second hypersurface ${}^{(3)}\mathcal{S}''$ be described at small distances by giving t as some reasonable and regular function t" of r, θ, and φ, going over at large distances into the formula

$$t'' = (8m_1^*\, r)^{\frac{1}{2}} \qquad (62)$$

with m_1^* = a constant. Taking the differential of this expression and substituting into (58), one finds that the second hypersurface has the asymptotic geometry

$$ds^2 \sim [1 + 2(m^* - m_1)/r]dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\varphi^2) \qquad (63)$$

The masses not only can be different but -- in the example -- must be different! One's first surprise at this result traces back to a semantic obscurity in the word "flat":

Meaning 1: The intrinsic 3-geometry is asymptotically flat.

Meaning 2: The intrinsic 3-geometry is asymptotically flat and also the extrinsic curvature is zero.

Only when "flat" is used in sense 2 do the apparent masses have to agree between two asymptotically flat geometries. However, the two-surface formulation of relativity focusses on intrinsic 3-geometry, so that "flat" there is used in sense 1. There is no problem of compatibility between the two 3-geometries in the example. René Thom[25] has even shown that one can fill in between two 3-geometries of different topology with a non-singular topology. Whether and when the geometry laid down on that topology can also be non-singular is a deeper question!

Question of Effectively Elliptic Character of the Thin Sandwich Problem

Does the CIVP (58) -- or the triplet of differential equations to which it corresponds -- have elliptic character? This issue brings to mind the question whether the equation

$$d^2\Psi/d\theta^2 + (\lambda - V_o\cos\theta)\Psi = 0 \qquad (64)$$

has eigenvalue character. One might think not, to look at the regions of θ where the "oscillation factor" or "effective kinetic energy factor" $(\lambda - V_o\cos\theta)$ is negative. There the

(25) R. Thom, Comment. Math. Helv. 28, 17 (1954), Chapter IV. See also J. W. Smith, U. S. Nat. Acad. Sci., Proc. 46, 111 (1960)

solution is curved away from the θ axis. However, what counts in the end for the question of nodes and eigenvalues is the region where this factor is positive and the solution is oscillatory. The equation is <u>effectively</u> oscillatory in character (for λ sufficiently in excess of $-V_0$). It is difficult in the case of (58,59) to be precise at this stage; but one has the impression that it is in a comparable sense <u>effectively elliptic</u>. Space in the "thin sandwich" problem is divided up ordinarily into regions where $(2T^{*\perp\perp} - {}^{(3)}R)$ is positive -- and where therefore also the shift anomaly γ_2 has to be positive -- and regions where the second quantity has to follow the first in changing sign. At the interface between one such region and another the anomaly γ_2 has to change sign. This situation reminds one -- to use another analogy -- of the theory of buckling of shells, and of conditions at the boundary between one region of crumpling and another.

As the shift anomaly γ_2 now comes so centrally into the discussion, a few words about it are in order. Consider the equation for the eigenvalues of the extrinsic curvature tensor K_{ik} -- or rather, of the closely related shift tensor $\gamma_{ik} = N_0 K_{ik}$. Consider the determinant

$$\begin{vmatrix} (\gamma_1{}^1 - \lambda) & \gamma_1{}^2 & \gamma_1{}^3 \\ \gamma_2{}^1 & (\gamma_2{}^2 - \lambda) & \gamma_2{}^3 \\ \gamma_3{}^1 & \gamma_3{}^2 & (\gamma_3{}^3 - \lambda) \end{vmatrix}$$

$$= \det \gamma_i{}^k - (\gamma_2/2)\lambda + (\mathrm{Tr}\ \underset{\sim}{\gamma})\lambda^2 - \lambda^3 \qquad (65)$$

A change in coordinates changes the $\gamma_i{}^k$ individually but not the eigenvalues λ and consequently not the coefficients of the various powers of λ on the right hand side of (65). Therefore consider a system of coordinates such that at the particular point of interest the shift tensor $\gamma_i{}^k$ is diagonal. Let the elements down the diagonal -- the eigenvalues λ -- be denoted by A, B, C. Then the coefficient of $-\lambda$ in the expansion of the secular determinant $(A-\lambda)(B-\lambda)(C-\lambda)$ is

$$(BC + CA + AB) = \frac{1}{2}\left[(A + B + C)^2 - (A^2 + B^2 + C^2)\right]$$

$$= \frac{1}{2}[(\mathrm{Tr}\ \underset{\sim}{\gamma})^2 - \mathrm{Tr}\ \underset{\sim}{\gamma}^2] = (1/2)(\text{shift anomaly}) = \gamma_2/2$$

Associated with the point in question consider a three dimensional space with coordinates A, B, C. Then the shift tensor is represented by a single point in this space. Moreover this point is independent of the choice of coordinate system in the hypersurface. In the space (A, B, C) construct through the origin a line with direction cosines $(3^{-\frac{1}{2}}, 3^{-\frac{1}{2}}, 3^{-\frac{1}{2}})$. Construct a double cone with this line as axis with an angle of opening θ such that

$$\cos\theta = 3^{-\frac{1}{2}} = \text{scalar product of } 3^{-\frac{1}{2}}(1,1,1) \text{ with } \left\{\begin{matrix} (1,0,0) \text{ or} \\ (0,1,0) \text{ or} \\ (0,0,1) \end{matrix}\right\} \quad . \tag{67}$$

Then any point <u>on</u> a coordinate axis lies <u>on</u> one or other half of the cone. Every point on a coordinate axis also annuls the shift anomaly, according to (66). It takes only a few more steps to show that the shift anomaly γ_2 is

(1) zero for <u>every</u> point <u>on</u> either cone;

(2) positive for every point <u>within</u> either cone; and

(3) negative in the neutral space <u>between</u> cones.

To each of these three cases may be said to correspond a particular character of th shift tensor $\gamma_i{}^k$. What is the detailed value of the shift tensor is only settled by extremization of the CIVP -- or by integration of the initial value equations with appropriate boundary condition -- and is therefore governed by the initial value data all over the hypersurface. However, only the <u>local</u> value of the quantity $(2T^{**11} - {}^{(3)}R)$ -- read out of initial value data -- is required to determine the <u>character</u> of the shift tensor. Turn now from comments on the general problem to a particular example.

<u>Example Where Both Hypersurfaces that Bound the Thin Sandwich have Ideal 3-Sphere Geometry</u>

Let both hypersurfaces have the geometry of the ideal sphere

$$x^2 + y^2 + z^2 + w^2 = 1;$$

thus for ${}^{(3)}\mathscr{S}'$ (give it the <u>name</u> x^o !)

$$ds^2 = a'^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\varphi^2)] \qquad (68)$$

and for ${}^{(3)}\mathscr{S}''$ (give it the <u>name</u> $x^o + \Delta x^o$!)

$$ds^2 = a''^2[d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\varphi^2)] \qquad (69)$$

where a' and a" are constants. Or to use another language, consider a one parameter family of such hypersurfaces, characterized by a parameter x^o:

$$a = a(x^o) \qquad (70)$$

and pick some fixed value of x^o, thus specifying

$$a \quad \text{and} \quad da/dx^o \ (\sim(a'' - a')/\Delta x^o). \qquad (71)$$

(As remarked earlier, the value of Δx^o will drop out of the results at the end.) The remaining initial value data comprises the energy flow, which we set equal to zero, and the energy density, which we assume independent of position:

$$T^{**\perp\perp} = \text{constant (independent of } \chi, \theta, \varphi) \qquad (72)$$

The question now is: What 4-geometry to fill in between the two hypersurfaces so as to satisfy the thin sandwich equations? The time-like perpendicular erected to ${}^{(3)}\mathscr{S}'$ at the point

χ, θ, φ will have to be assigned a certain length. Also it will be necessary to tell what point it touches on the hypersurface ${}^{(3)}\mathcal{S}''$, or to tell what the starred quantities are in the following formula for the coordinates of this point:

$$\chi - \chi^*, \quad \theta - \theta^*, \quad \varphi - \varphi^* \tag{73}$$

On account of the symmetry of the sphere it will be simplest to assume -- as a trial -- the same angles for both points, or to take all the starred quantities equal to zero. Thus the shift function is assumed zero:

$$N^{\chi} = \chi^*/\Delta x^o = 0, \text{ etc. (three equations)} \tag{74}$$

Now for the shift tensor! It has to do with the fractional increase -- between one hypersurface and the other -- in the distance between points with corresponding coordinates, say (χ, θ, φ) and $(\chi + d\chi,\ \theta + d\theta,\ \varphi + d\varphi)$. But this increase for the case we are considering is the same in all directions and at all places, and is in direct proportion to the fractional increase in the value of the radius. Thus the eigenvalues of $\gamma_i{}^k$ are identical:

$$A = B = C = \frac{\text{(fractional increase in radius)}}{\text{(change in the highly nominal parameter } x^o)}$$

$$= (1/a)(da/dx^o) \tag{75}$$

The point in the space (A,B,C) lies inside one half of the double cone, right on the axis. The shift anomaly is positive:

$$\gamma_2 = (\mathrm{Tr}\ \underset{\sim}{\gamma})^2 - \mathrm{Tr}\ \underset{\sim}{\gamma}^2 = (6/a^2)(da/dx^o)^2, \tag{76}$$

but independent of position. Likewise the covariant derivative of $\gamma_i^{\ k}$ is zero and the N_i vanish. These circumstances guarantee that the condensed initial value equations (59) are automatically satisfied. It only remains to find the lapse function N_o:

$$\gamma_2/N_o^2 = 2T^{**\perp\perp} - {}^{(3)}R \tag{77}$$

or

$$(6/a^2)(da/N_o dx^o)^2 = 2T^{**\perp\perp} - 6/a^2. \tag{78}$$

Instead of actually solving for N_o, it is better to recognize that $N_o dx^o$ is the <u>proper</u> time separation -- call it dt -- between hypersurfaces, the parameters attached to which are x^o and $x^o + dx^o$, and is therefore directly the physical quantity of interest. Thus write

$$(da/dt)^2 = (a^2/3)T^{**} - 1. \tag{79}$$

The dynamics of the model universe are completely determined by (79) as soon as one puts in the law of change of energy density with expansion:

$$T^{**\perp\perp} = (8\pi G/c^4)(Mc^2/2\pi^2 a^3) \tag{80}$$

for a universe filled with inchoate dust (Friedmann universe); and

$$T^{**\perp\perp} = \mathrm{constant}/a^4 \tag{81}$$

for a system filled with isotropic radiation (Tolman universe).

Question of Uniqueness. The Linear Approximation

The purpose here was not to take up old problems anew, but to prepare the way in a simple example to investigate the uniqueness of the 4-geometry determined by ${}^{(3)}g_{ik}$, $\partial\,{}^{(3)}g_{ik}/\partial x^o$, $T^{*\perp i}$ and $T^{*\perp\perp}$. Suppose the vector shift function $N^i = (\chi^*, \theta^*, \varphi^*)/\Delta x^o$ is not assumed to be zero but investigated in terms of the equations themselves. Will one find oneself with no alternative except the familiar solution already sketched out? Unfortunately the three coupled second order equations to be solved are only quasilinear, not linear. The problem appears difficult without some deeper mathematical considerations to draw on which do not present themselves immediately. Therefore no decisive results can be offered here. What has been investigated is the case where the contribution of the shift vector N_i to the shift tensor

$$\gamma_{ik} = (1/2)(N_{i|k} + N_{k|i} - \partial\,{}^{(3)}g_{ik}/\partial x^o) \tag{82}$$

is so small compared to the "main term" (Eq. 75) that one is justified in treating the condensed initial value eqs. (59) as linear in N^i. These equations then take the form

$$(\sin\chi)^{-2}(\partial^2\chi^*/\partial\theta^2 + (\sin\theta)^{-2}\partial^2\chi^*/\partial\varphi^2 + \cot\theta\,\partial\chi^*/\partial\theta$$
$$+ 4\chi^*) - (\partial/\partial\chi)(\partial\varphi^*/\partial\varphi + \sin\theta)^{-1}(\partial/\partial\theta)(\theta^*\sin\theta) = 0, \tag{83}$$

and

$$\sin\chi(\partial/\partial\chi)(\sin\chi\ \partial\theta^*/\partial\chi)+(\sin\theta)^{-2}(\partial^2\theta^*/\partial\varphi^2)$$
$$+\ 2\theta^*\ -(\sin\chi)^{-3}(\partial/\partial\chi)(\sin^3\chi\ \partial\chi^*/\partial\theta)$$
$$-(\sin\theta)^{-2}(\partial/\partial\theta)(\sin^2\theta\ \partial\varphi^*/\partial\varphi)\ =\ 0 \tag{84}$$

and

$$\sin^2\theta\ \sin\chi(\partial/\partial\chi)(\sin\chi\ \partial\varphi^*/\partial\chi)+\sin\theta(\partial/\partial\theta)(\sin\theta\ \partial\varphi^*/\partial\theta)$$
$$-(\sin\chi)^{-3}(\partial/\partial\chi)(\sin^3\chi\ \partial\chi^*/\partial\varphi)-\sin\theta(\partial/\partial\theta)(\sin\theta\ \partial\theta^*/\partial\varphi)\ =\ 0 \tag{85}$$

One can seek a solution by writing

$$\chi^*(\chi,\theta,\varphi)\ =\ \Sigma f_{\ell,m}(\chi)\ Y_\ell^{(m)}(\theta,\varphi) \tag{86}$$

No thoroughgoing analysis along this line has been completed. However, Professor C. W. Misner was kind enough to point out at the Warsaw conference that the equations ought in principle to admit of <u>rotations</u>. This point has since been tested and verified. It obviously makes no difference to the geometry of the 3-sphere ${}^{(3)}\mathcal{G}''$ whether one set of hyperspherical polar coordinates χ,θ,φ or a rotated set is used to describe the location of the points. However, it does make a difference to the coordinate-dependent shift vector N^k. To fill in between ${}^{(3)}\mathcal{G}'$ and ${}^{(3)}\mathcal{G}''$ with a thin-sandwich ${}^{(4)}\mathcal{G}$ -- compatible with the intrasurface variational principle or initial value equations -- does not in itself fix the values of these quantities. The time-like normals that reach between the one hypersurface and the other, which start at (χ,θ,φ) on one hypersurface, and also end at $(\chi,\theta,\varphi,)$ on the other hypersurface, will end at <u>different</u> values of (χ,θ,φ) when a rotated coordinate system is used: $(\chi-\chi^*,\theta-\theta^*,\varphi-\varphi^*)$.

<u>Shifts Produced by the Six Independent Rotations</u>

The calculation of the starred changes in the angles under a typical small rotation is most easily made by going to cartesian coordinates:

$$
\begin{aligned}
x &= a \sin\chi \sin\theta \cos\varphi \\
y &= a \sin\chi \sin\theta \sin\varphi \\
z &= a \sin\chi \cos\theta \\
w &= a \cos\chi
\end{aligned}
\tag{87}
$$

There are six independent small rotations out of which the most general small rotation is constructed by linear combination. Consider as an example a turn by the small angle θ_{zw} in the (z,w) plane

$$
\begin{aligned}
dx &= 0, \qquad dy = 0 \\
dz &= \theta_{zw}w, \\
dw &= -\theta_{zw}z.
\end{aligned}
\tag{88}
$$

The resulting change in the polar angle θ is

$$
\begin{aligned}
d\theta &= \cos^2\theta \; d(\tan\theta) = \cos^2\theta \; d[(x^2+y^2)^{\frac{1}{2}}/z] \\
&= -\cos^2\theta \; (x^2+y^2)^{\frac{1}{2}} \; x^{-2} \; \theta_{zw}w \\
&= -\cot\chi \sin\theta \; \theta_{zw}
\end{aligned}
\tag{89}
$$

Similarly one finds the changes in all three coordinate angles under all six independent rotations (Table V).

Table V. Changes in polar angles on 3-sphere brought about by the six independent types of rotation.

	χ^*	θ^*	φ^*
θ_{yz}	0	$\sin\varphi$	$\cot\theta\cos\varphi$
θ_{zx}	0	$-\cos\varphi$	$\cot\theta\sin\varphi$
θ_{xy}	0	0	-1
θ_{xw}	$\sin\theta\cos\varphi$	$\cot\chi\cos\theta\cos\varphi$	$-\cot\chi\sin\varphi/\sin\theta$
θ_{yw}	$\sin\theta\sin\varphi$	$\cot\chi\cos\theta\sin\varphi$	$\cot\chi\cos\varphi/\sin\theta$
θ_{zw}	$\cos\theta$	$-\cot\chi\sin\theta$	0

It is easy to verify that each line of Table V represents a solution of the linearized initial value Eqs. (83,84,85). It is the conjecture that there are no other independent solutions of these equations which are free of truly geometrical singularity -- as distinguished from coordinate singularity[25] over the entire 3-sphere.

Even if and when this conjecture can be established, there will remain the question of uniqueness of the equations for this

(25) In principle all question of what is a coordinate singularity and what is a truly geometrical singularity can and should be eliminated by the use of two or more coordinate patches (reference 18) to eliminate all singularities in the coordinate systems that cover the 3-sphere.

two sphere problem in their full non-linear form (59). After that will be the question of uniqueness in more general situations.

Assessment of Mach's Principle

Pending the investigation of these apparently difficult mathematical questions, it would not appear unreasonable to adopt as a working hypothesis the position (formulation 4 of Mach's principle) that the specification of a sufficiently regular closed 3-dimensional geometry at two immediately succeeding instant , and of the density and flow of mass-energy, is to determine the geometry of spacetime, past, present and future, and thereby the inertial properties of every infinitesimal test particle. In this sense it is proposed to view Mach's principle as the boundary condition for Einstein's field equations, and an essential part of the "plan" of general relativity. The condensed intrasurface variational principle (58) is the most compact mathematical statement available of this interpretation of Mach's principle. As conceived here, it carries with it the tacit requirement that the model universe be closed.

ACKNOWLEDGMENT

Thanks are expressed to R.F. Baierlein, H. Bondi, R.H. Dicke, E. Guth, J.E. Hogarth, A. Komar, L. Markus, C.W. Misner, E. Schucking, D. Sciama, D.H. Sharp, A.H. Taub, R. Utiyama, E.P. Wigner and other colleagues for discussions; to Professor Hideki Yukawa and Professor Minoru Kobayasi of Kyoto University and to Professor Wesley Brittin of the University of Colorado for the hospitality of the two institutions where this report was written; to the organizers of the Warsaw conference for their friendly helpfulness and the stimulus of the meeting they arranged; and to the Office of Scientific Research of the U.S. Air Force for a contribution towards the cost of travel to the meeting.

APPENDIX: THE TAUB UNIVERSE INTERPRETED IN TERMS OF GRAVITATIONAL RADIATION OF MAXIMAL WAVE LENGTH.

The Taub universe[8] is free of any "real matter" at all. Taub derived this solution of Einstein's equations,

$$d\sigma^2 = -d\tau^2 = \gamma_1 \; dx^2 + (\gamma_1 \sin^2 x + \gamma_3 \cos^2 x) dy^2$$

$$+ 2\gamma_3 \cos x \; dy \; dz + \gamma_3 dz^2 - \gamma_1{}^2 \; \gamma_3 dt^2 \; ,$$

with

$$\gamma_1 = \cosh t/4 \cosh^2(t/2),$$

$$\gamma_3 = 1/\cosh t,$$

from arguments of group theory having nothing directly to do with the kind of considerations which are the center of attention in this report. Therefore, it is of interest to see how one can be headed towards the same solutions by a natural physical line of reasoning.

Replace the dust in the Friedman universe by electromagnetic radiation distributed uniformly in space and in direction. One arrives at the Tolman universe[26]. During its expansion and recontraction the wave length of every standing wave varies as the radius a of the model universe.

---

[26] R.C. Tolman, <u>Relativity, Thermodynamics and Cosmology</u>, Clarendon Press, Oxford, 1934.

In consequence the density of mass-energy varies not as $1/a^3$, as in the Friedmann universe, but as $1/a^4$. Replace the electromagnetic radiation by gravitational radiation of short wave length. There isn longer any "real" density of mass-energy on the right hand side of Einstein's equation. However, the fine scale ripples in the geometry bring about the same type of larger scale curvature as would be caused by a "real" distribution of mass-energy. Let δg denote the local root mean square amplitude of the fluctuations in the metric and let $ƛ = \lambda/2\pi$ = (wave length)/2π denote their reduced wave length. Then the effective density of mass-energy associated with the gravitational radiation is of the order

$$\hat{T}_{\perp\perp\ \text{effective}} \quad (c^4/8\pi G)(\delta g/ƛ)^2$$

To curve a space up into closure with a radius which at the moment of maximum expansion has the value a_o requires an energy density given by the equation

$$^{(3)}R = (16\pi G/c^4)\hat{T}_{\perp\perp} \quad ,$$

or

$$6/a_o^2 \sim 2(\delta g/ƛ)^2 \quad .$$

Thus the amplitude of the ripples need not be great

$$\delta g(\text{at maximum expansion}) \quad 3^{1/2}\, ƛ/a_o$$

if the wave length is short.

During the expansion and recontraction the energy density, proportional to $(\delta g/ƛ)^2$, necessarily varies as $1/a^4$.

Consequently the amplitude of the ripples varies in accordance with the formula

$$\begin{aligned}\delta g(t) &\sim \text{constant}_1\ \chi(t)/a^2(t)\\ &\sim \text{constant}_2/a(t)\\ &\sim 3^{1/2}\ \chi_0/a(t)\\ &\sim (3^{1/2}/n)(a_0/a(t))\end{aligned}$$

Here the last expression refers to the case where the perturbation in the otherwise ideal spherical geometry is described by a hyperspherical harmonic[27] of order n. It is not reasonable to consider the factor $3^{1/2}$ in this order of magnitude formula as a reliable number!

From considering a gravitational wave of very short wave length it is natural to turn to the opposite limiting case where the order n has the minimum possible value and the wave length has the maximum possible value which will fit into the 3-sphere. The corresponding hyperspherical harmonic has well defined symmetry properties[27]. Possession of these symmetry properties, and of the critical amplitude required for closure, are the features of the special <u>gravitational wave that gives the Taub universe</u>.

The Taub universe is homogeneous but not isotropic: the curvature differs from one direction to another, but the principal values of the curvature do not change from place to place.

---

(27) E. Lifshitz, J. Phys. U.S.S.R. <u>10</u>, 116 (1946).

The curvature provides a more reasonable way of talking about the perturbations in the geometry than does the quantity δg for a well known reason: neither out of the metric coefficients nor out of their first derivatives can one form coordinate-~~independent~~ independent quantities. For the order of magnitude of typical components of the fluctuation part of the curvature in a local Lorentz frame one has the estimate

$$
\begin{aligned}
\hat{R}(t)_{\text{wave}} &\sim \delta g/\bar{\lambda}^2 \\
&\sim \delta g/(a/n)^2 \\
&\sim n a_o/a^3(t) \quad ,
\end{aligned}
$$

as compared to the typical component of the curvature of the background geometry,

$$\hat{R}_{\text{background}} \sim 1/a^2(t) \quad .$$

Thus the mode of longest wave length and lowest n is the one for which the perturbations in the geometry -- as measured by the differences in the curvature in different directions -- are not greatest (as one might have thought from the expression for δg) but least.

At early and late stages this perturbation becomes percentagewise larger and larger and the geometry ultimately develops infinite curvature, in accordance with what appears to be a general principle[28].

---

[28] GMD pp.61-64.